



An International ICT Journal
Featuring Industry-University-Institute Cooperation

ISSN 1673-5188
CN 34-1294/TN

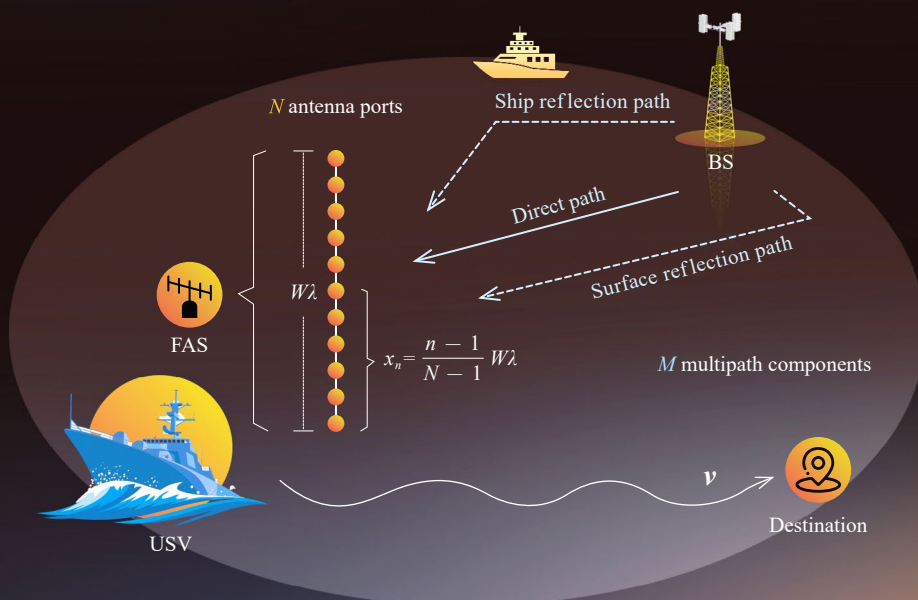
ZTE COMMUNICATIONS

中兴通讯技术(英文版)

<https://zte.magtechjournal.com>

September 2026, Vol. 24 No. 3

Special Topic: Reconfigurable Antenna Systems for Next-Generation Mobile Communications



(See Fig. 1 on P. 20)

ISSN 1673-5188



The 10th Editorial Board of ZTE Communications

Chairman

Gao Wen, Peking University (China)

Vice Chairmen

Xu Ziyang, ZTE Corporation (China) | **Xu Chengzhong**, University of Macau (China)

Members (Surname in Alphabetical Order)

Ai Bo	Beijing Jiaotong University (China)
Cao Jiannong	The Hong Kong Polytechnic University (China)
Chen Chang Wen	The Hong Kong Polytechnic University (China)
Chen Yan	Northwestern University (USA)
Chi Nan	Fudan University (China)
Cui Shuguang	UC Davis (USA) and The Chinese University of Hong Kong, Shenzhen (China)
Fang Rong	ZTE Corporation (China)
Gao Wen	Peking University (China)
Gao Yang	Nanjing University (China)
Gao Yue	Fudan University (China)
Ge Xiaohu	Huazhong University of Science and Technology (China)
Guo Yike	The Hong Kong University of Science and Technology (China)
He Yejun	Shenzhen University (China)
Victor C. M. Leung	The University of British Columbia (Canada)
Li Xiangyang	University of Science and Technology of China (China)
Liao Yong	Chongqing University (China)
Lin Xiaodong	ZTE Corporation (China)
Liu Chi	Beijing Institute of Technology (China)
Liu Jian	ZTE Corporation (China)
Liu Yue	Beijing Institute of Technology (China)
Ma Jianhua	Hosei University (Japan)
Ma Zheng	Southwest Jiaotong University (China)
Pan Yi	Shenzhen University of Advanced Technology, Chinese Academy of Sciences (China)
Peng Mugen	Beijing University of Posts and Telecommunications (China)
Ren Fuji	Tokushima University (Japan)
Ren Kui	Zhejiang University (China)
Sheng Min	Xidian University (China)
Su Zhou	Xi'an Jiaotong University (China)
Sun Huifang	Pengcheng Laboratory (China)
Sun Zhili	University of Surrey (UK)
Tao Meixia	Shanghai Jiao Tong University (China)
Wang Chengxiang	Southeast University (China)
Wang Haiming	Southeast University (China)
Wang Ling	Northwestern Polytechnical University (China)
Wang Xiang	ZTE Corporation (China)
Wang Xiyu	ZTE Corporation (China)
Wang Yongjin	Nanjing University of Posts and Telecommunications (China)
Xu Chengzhong	University of Macau (China)
Xu Ziyang	ZTE Corporation (China)
Yang Kun	University of Essex (UK)
Yu Hongfang	University of Electronic Science and Technology of China (China)
Yu Zhiwen	Harbin Engineering University (China)
Yuan Jinhong	University of New South Wales (Australia)
Zeng Wenjun	Eastern Institute of Technology, Ningbo (China)
Zhang Honggang	Macau University of Science and Technology (China)
Zhang Jianhua	Beijing University of Posts and Telecommunications (China)
Zhang Rui	National University of Singapore (Singapore)
Zhang Wenjun	Shanghai Jiao Tong University (China)
Zhang Wenqiang	Fudan University (China)
Zhang Yueping	Nanyang Technological University (Singapore)
Zhou Wanlei	City University of Macau (China)
Zhuang Weihua	University of Waterloo (Canada)

CONTENTS

ZTE COMMUNICATIONS
September 2026 Vol. 24 No. 3 (Issue 96)

Special Topic ►

Reconfigurable Antenna Systems for Next-Generation Mobile Communications

- 01 Editorial Yang Kun, Song Lingyang, Wong Kai-Kit, Zhao Yizhe
- 03 Closed-Form Outage Expression of Fluid Antenna Systems for Ultra-Reliable Communications
..... Wu Ruojia, Xu Rui, Ye Yinghui, Song Xi, Lu Guangyue
- 12 Equivalent Degrees of Freedom Based Secrecy Outage Analysis for Fluid Antenna Wiretap
Channels..... Wu Tuo, Wu Zhenguan, Zhi Kangda, Liu Baiyang, Tong Kin-Fai
- 18 Prediction-Window Aided Fluid Antenna Systems for High-Mobility Communications
..... Huang Yinning, Zeng Cheng, Yang Zhaohui
- 31 Joint Resource Allocation for Movable-Antenna-Assisted Secure Integrated Data and Energy
Transfer..... Xiang Pengrong, Li Zhiguang, Zhao Yizhe
- 40 Robust Movable Antenna Position Optimization Under Imperfect Channel Path Parameters
..... Jia Qiaoran, Mei Weidong, Wei Xin, Chen Zhi, Ning Boyu
- 48 High Resolution Automotive Radar Imaging Integrating Doppler Division Multiple Access and
Frequency Agility Zhang Wei, Zhang Chenyong, Wang Yu, Ding Jin, Li Yao
- 59 Constructing Large-Scale Real-World Benchmark Datasets for AIOps
..... Li Zeyan, Han Jing, Zhao Nengwen, Zhang Shenglin, Sun Yongqian, Chen Pengfei,
Wen Xidao, Ma Minghua, Pei Dan
- 70 Reconfigurable Metasurface for Near-Field Efficient Wireless Power Transfer System
..... Wang Yunfeng, Xu Zhiyu, Zhang Hongtao
- 82 An Embedded Database System for YANG
..... Han Bing, Gao Jirun, Zhang Kaiqi, Wang Hongzhi, Chen Hedui
- 92 A Method for Text Information Extraction Leveraging Large Language Models and Two-Stage In-
formation Filtering Lu Ping, Qin Bin, Wang Yiyang, Xu Yan, Deng Fangwei
- 11 New Member (Zhang Wenjun) of *ZTE Communications* Editorial Board

Technology ► Frontiers

Industry-Academia ► Co-Research

Roundup ►

Serial parameters: CN 34-1294/TN*2003*q*16*98*en*P*¥30.00*2200*11* 2026-09

Special Topics for 2026

Special Topic	Leading Guest Editor
1 Achievements of ZTE's Industry-University-Institute Cooperation Projects	Xu Chengzhong (University of Macau, China)
2 AI-Agent Communication Network (ACN): Architecture, Protocols and Key Technologies	Sun Tao (China Mobile Research Institute, China)
3 Reconfigurable Antenna Systems for Next-Generation Mobile Communications	Yang Kun (Nanjing University, China)
4 Goal-Oriented Joint Semantic and Channel Coding for Future Communications	Yuan Jinhong (University of New South Wales, Australia)

ZTE Communications Guidelines for Authors

Remit of Journal

ZTE Communications publishes original theoretical papers, research findings, and surveys on a broad range of communications topics, including communications and information system design, optical fiber and electro-optical engineering, microwave technology, radio wave propagation, antenna engineering, electromagnetics, signal and image processing, and power engineering. The journal is designed to be an integrated forum for university academics and industry researchers from around the world.

Manuscript Preparation

Manuscripts must be typed in English and submitted electronically in MS Word (or compatible) format. The word length is approximately 3 000 to 8 000, and no more than 8 figures or tables should be included. Authors are requested to submit mathematical material and graphics in an editable format.

Abstract and Keywords

Each manuscript must include an abstract of approximately 150 words written as a single paragraph. The abstract should not include mathematics or references and should not be repeated verbatim in the introduction. The abstract should be a self-contained overview of the aims, methods, experimental results, and significance of research outlined in the paper. Three to eight carefully chosen keywords must be provided with the abstract.

References

Manuscripts must be referenced at a level that conforms to international academic standards. All references must be numbered sequentially in-text and listed in corresponding order at the end of the paper. References that are not cited in-text should not be included in the reference list. References must be complete and formatted according to *ZTE Communications* Editorial Style. A minimum of 10 references should be provided. Footnotes should be avoided or kept to a minimum.

Content and Structure

ZTE Communications seeks to publish original content that may build on existing literature in any field of communications. Authors should not dedicate a disproportionate amount of a paper to fundamental background, historical overviews, or chronologies that may be sufficiently dealt with by references. Authors are also requested to avoid the overuse of bullet points when structuring papers. The conclusion should include a commentary on the significance/future implications of the research as well as an overview of the material presented.

Peer Review and Editing

All manuscripts will be subject to a two-stage anonymous peer review as well as copyediting, and formatting. Authors may be asked to revise parts of a manuscript prior to publication.

Biographical Information

All authors are requested to provide a brief biography (approx. 100 words) that includes email address, educational background, career experience, research interests, awards, and publications.

Acknowledgments and Funding

A manuscript based on funded research must clearly state the program name, funding body, and grant number. Individuals who contributed to the manuscript should be acknowledged in a brief statement.

Address for Submission

<http://mc03.manuscriptcentral.com/ztecom>

Submission of a manuscript implies that the submitted work has not been published before (except as part of a thesis or lecture note or report or in the form of an abstract); that it is not under consideration for publication elsewhere; that its publication has been approved by all co-authors as well as by the authorities at the institute where the work has been carried out; that, if and when the manuscript is accepted for publication, the authors hand over the transferable copyrights of the accepted manuscript to *ZTE Communications*; and that the manuscript or parts thereof will not be published elsewhere in any language without the consent of the copyright holder. Copyrights include, without spatial or timely limitation, the mechanical, electronic and visual reproduction and distribution; electronic storage and retrieval; and all other forms of electronic publication or any other types of publication including all subsidiary rights.

Responsibility for content rests on authors of signed articles and not on the editorial board of *ZTE Communications* or its sponsors.

Statement

This magazine is a free publication for you. If you do not want to receive it in the future, you can send the "TD unsubscribe" mail to magazine@zte.com.cn. We will not send you this magazine again after receiving your email. Thank you for your support.



Special Topic on Reconfigurable Antenna Systems for Next-Generation Mobile Communications

Guest Editors



Yang Kun



Song Lingyang



Wong Kai-Kit



Zhao Yizhe

Driven by the rapid development of 6G and beyond mobile communication technologies, future wireless networks are continuously evolving toward ultra-high data rates, ultra-reliable transmission, and massive ubiquitous connectivity, which significantly accelerates the iterative upgrading and intelligent transformation of modern wireless communication systems. As a core and enabling technology for enhancing network flexibility, environmental adaptability and anti-fading robustness, reconfigurable antenna systems have become indispensable to support the diversified and high-performance requirements of next-generation wireless communications. Representative and emerging research hotspots in this field include fluid antennas, movable antennas, pinching antennas, and reconfigurable holographic antennas. Fluid antennas utilize flexible liquid conductive materials to dynamically reshape antenna structures and adjust key electromagnetic parameters, exhibiting promising application prospects in high-frequency millimeter-wave and terahertz communication scenarios. Movable antennas effectively optimize wireless signal propagation paths, transmission quality and system throughput by flexibly adjusting their spatial positions and radiation attitudes, with intelligent resource optimization and mechanical

stability design serving as key research directions. Pinching antennas achieve flexible electromagnetic reconfiguration through controllable mechanical clamping and deformation, making them highly suitable for lightweight, flexible wearable devices and miniaturized Internet of Things (IoT) terminals. In addition, reconfigurable holographic antennas realize precise and adaptive beamforming based on electromagnetic holographic principles, and can be well integrated with advanced multiple-input multiple-output (MIMO) and non-orthogonal multiple access (NOMA) technologies to further improve the spectral efficiency, interference suppression capability, and communication reliability of wireless systems. It is worth noting that the research scope of reconfigurable antenna technologies is extremely broad, and the above architectures only represent a part of current research advances, rather than covering all innovative antenna solutions for future mobile networks.

In this special issue, we aimed to collect state-of-the-art studies focusing on the design, analysis and optimization of these innovative reconfigurable antenna systems for next-generation mobile communications. We sincerely welcomed high-quality original research and review contributions that investigate the theoretical foundations, practical design schemes, performance evaluation methodologies, and real-world application scenarios of advanced reconfigurable antenna systems tailored for future mobile communication networks. Our open call for papers successfully attracted numerous valuable submissions from researchers worldwide, fully demonstrating the surging research attention and vibrant development momentum within this fast-evolving wireless domain. After two rigorous and intensive rounds of anonymous peer review, we have carefully selected

DOI: 10.12142/ZTECOM.202603001

Citation (Format 1): Yang K, Song L Y, Wong K K, et al. Editorial: reconfigurable antenna systems for next-generation mobile communications [J]. *ZTE Communications*, 2026, 24(3): 1 – 2. DOI: 10.12142/ZTECOM.202603001

Citation (Format 2): K. Yang, L. Y. Song, K. K. Wong, et al., "Editorial: reconfigurable antenna systems for next-generation mobile communications," *ZTE Communications*, vol. 24, no. 3, pp. 1–2, Sept. 2026. doi: 10.12142/ZTECOM.202603001.

and curated a set of insightful, well-refined papers with solid innovation and practical significance, which are finally featured in this special issue.

The first paper, titled “Closed-Form Outage Expression of Fluid Antenna Systems for Ultra-Reliable Communications”, proposes an adjustable-precision approximation for ultra-reliable fluid antenna systems (FAS). Conventional FAS outage formulas either incur heavy computation or sacrifice accuracy. It adopts a block-correlated Rayleigh fading model and an N-term series expansion for the correlated-fading Bessel function, using N to balance precision and complexity. This method offers an efficient analytical tool for FAS outage evaluation and guides practical ultra-reliable FAS design.

The second paper, titled “Equivalent Degrees of Freedom Based Secrecy Outage Analysis for Fluid Antenna Wiretap Channels”, adopts equivalent degrees of freedom (EDoF) to build a closed-form secrecy outage framework under the Jakes model. It maps correlated physical ports to a small set of dominant spatial modes with eigenvalue weighting, simplifying wiretap-channel analysis for Bob and Eve. Derived theoretical results quantify secrecy effects of aperture sizes and signal-to-noise ratio (SNR) scaling. This method guides aperture planning and physical-layer security optimization for fluid antennas.

The third paper, titled “Prediction-Window Aided Fluid Antenna Systems for High-Mobility Communications”, resolves the scanning overhead bottleneck of fluid antennas for high-mobility intelligent agents via a prediction-tolerant window scheme. Centered on predicted antenna ports, the local sounding window leverages spatial correlation to tolerate prediction errors and restrain pilot overhead. The work builds a throughput optimization model and develops a lightweight two-stage algorithm for fast window tuning. This strategy cuts scanning overhead and improves throughput for fluid antennas supporting high-speed terminals.

The fourth paper, titled “Joint Resource Allocation for Movable-Antenna-Assisted Secure Integrated Data and Energy Transfer”, investigates movable-antenna supported secure IDET systems with passive eavesdroppers. Using a field-response channel model, it jointly optimizes beamformers and antenna positions under practical constraints to maximize the minimum secrecy rate, with energy signals reused as artificial noise. An iterative algorithm is designed to handle the resulting nonconvex problem. This framework supports resource optimization for secure movable-antenna data-and-energy transmission networks.

The fifth paper, titled “Robust Movable Antenna Position Optimization Under Imperfect Channel Path Parameters”, investigates robust beamforming and position optimization for downlink multiple-input single-output (MISO) movable antenna systems with imperfect channel information. Instead of conventional channel state information error modeling, it introduces bounded uncertainties for practical multipath parameters. By adopting discrete sampling and harmonic-mean approximation,

the work optimizes the worst-case SNR and develops an alternating optimization algorithm for beamforming and antenna position adjustment. This scheme provides a reliable robust optimization solution for movable antennas under inaccurate channel conditions.

To conclude, this special issue strives to advance the integration of reconfigurable antenna system designs for 6G and future wireless networks. We sincerely appreciate all authors for their original contributions and rigorous efforts, as well as the reviewers for their professional assessment and insightful suggestions. It is anticipated that the diverse research outcomes presented herein will offer valuable references, spark follow-up explorations, and foster continuous technical progress in the thriving field of reconfigurable antenna-enabled wireless communication systems.

Biographies

Yang Kun is a Junshi Distinguished Chair Professor with the School of Intelligent Software and Engineering and the Director of the Institute of Intelligent Networks and Communications (NINE), Nanjing University, China. His research interests mainly cover wireless communications and networks, and artificial intelligence empowered wireless systems. He has authored over 500 peer-reviewed papers and holds 50 granted patents. Since 2019, he has been a judge for the GSMA GLOMO Awards at MWC Barcelona. He received the 2024 IET Achievement Medal and the 2024 IEEE CommSoft TC Technical Achievement Award. He is a member of the Academia Europaea (MAE), an IEEE Fellow, an IET Fellow, and a Distinguished Member of the ACM.

Song Lingyang is a Boya Distinguished Professor at Peking University, China. His research focuses on cooperative wireless communications and networking. He is an IEEE Fellow and a Clarivate Analytics Highly Cited Researcher. He has received 27 IEEE Best Paper Awards, including the IEEE ComSoc Leonard G. Abraham Prize, as well as major honors such as the second prize of the National Natural Science Award and the first prize of the Natural Science Award from the Ministry of Education. He serves as a Standing Director of the China Institute of Communications, a member of the Board of Governors of the IEEE Vehicular Technology Society, and an Area Editor for *IEEE Transactions on Vehicular Technology*.

Wong Kai-Kit is a Chair Professor of wireless communications at the Department of Electronic and Electrical Engineering, University College London, UK. He previously held appointments at the University of Hong Kong, China, and the University of Hull, UK, as well as visiting positions at Lucent Technologies, Bell Labs and Stanford University, USA. His current research centers on 6G and beyond mobile communications. He is one of the early researchers who proposed multiuser MIMO. His first paper on multiuser MIMO, published at WCNC 2000, is considered the first research paper on this topic. He is a Fellow of the CIC, IEEE and IET. He served as the Editor-in-Chief for *IEEE Wireless Communications Letters* between 2020 and 2023.

Zhao Yizhe is currently an associate professor at the School of Information and Communication Engineering, University of Electronic Science and Technology of China. He was selected for the Youth Talent Support Project of the China Association for Science and Technology. His research achievement was selected as one of the Top 10 Scientific and Technological Advances in the Internet of Things Sector in China (2025). His research interests include simultaneous wireless information and power transfer, and reconfigurable antenna systems.



Closed-Form Outage Expression of Fluid Antenna Systems for Ultra-Reliable Communications

Wu Ruojia¹, Xu Rui¹, Ye Yinghui¹, Song Xi²,
Lu Guangyue¹

(1. Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
2. North China Electric Power University, Beijing 102206, China)

DOI: 10.12142/ZTECOM.202603002

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260828.1535.002.html>,
published online August 28, 2026

Manuscript received: 2026-01-07

Abstract: Due to its ability to significantly enhance communication reliability, the fluid antenna system (FAS) has emerged as a promising technology for ultra-reliable communications (URC). However, existing expressions to evaluate the outage probability of FAS are either computationally complex or simplified at the cost of accuracy. To address this issue, we propose an accuracy-controllable approximation method under Rayleigh fading channels. Specifically, we first approximate the fully correlated Rayleigh fading model as a block-correlation Rayleigh model, where the channel gains of ports in different blocks are assumed to be mutually independent, while the channel gains of ports in the same block are considered equally correlated, to simplify the analysis while maintaining accuracy. Then, we propose a novel method to approximate the Bessel function I_0 , introduced by the correlated Rayleigh fading, using an M -term series expansion, where the parameter M provides a tunable trade-off between approximation accuracy and computational complexity. Based on this approximation, a closed-form expression for the outage probability of FAS is derived. Simulation results demonstrate that our derived expression offers low computational complexity for evaluating the FAS outage probability while maintaining excellent approximation accuracy in the medium-to-high signal-to-noise ratio (SNR) range.

Keywords: ultra-reliable communication; fluid antenna system; approximation; outage probability

Citation (Format 1): Wu R J, Xu R, Ye Y H, et al. Closed-form outage expression of fluid antenna systems for ultra-reliable communications [J]. *ZTE Communications*, 2026, 24(3): 3 – 11. DOI: 10.12142/ZTECOM.202603002

Citation (Format 2): R. J. Wu, R. Xu, Y. H. Ye, et al., "Closed-form outage expression of fluid antenna systems for ultra-reliable communications," *ZTE Communications*, vol. 24, no. 3, pp. 3 – 11, Sept. 2026. doi: 10.12142/ZTECOM.202603002.

1 Introduction

Ultra-reliable communications (URC) is a key Beyond 5G (B5G) innovation that supports mission-critical applications, e. g. , autonomous vehicles and remote surgery, by ensuring over 99. 999% transmission reliability^[1-2]. Such ultra-reliability is often achieved through diversity techniques, particularly by realizing spatial diversity with multiple antennas in Multiple-Input Multiple-Output (MIMO) technology to combat signal fading^[3]. However, MIMO requires the minimum antenna spacing of half the wavelength to prevent electromagnetic coupling^[4]. This physical constraint makes it difficult to implement in highly miniaturized devices. To address this challenge, the fluid antenna system (FAS) has emerged. The core of FAS lies in its capability for dynamic reconfiguration. Specifically, this refers to a system

that can automatically switch the antenna position among N predefined locations, known as ports, within a given space^[5]. By selecting the port with the best channel condition, the system effectively leverages spatial diversity to maximize the received signal strength.

Accurate yet tractable performance analysis for FAS over Rayleigh fading is an important research direction, but it remains challenging due to the inherent spatial correlation among the fading channels. According to the classical Jakes' model, the spatial correlation between any two ports depends on the distance between the two ports Δd and the wavelength λ , and is characterized by the zero-order Bessel function of the first kind, i.e., $J_0(2\pi\Delta d/\lambda)$. Until now, considerable research efforts have been devoted to addressing this issue.

The authors in Ref. [6] proposed a reference point model where the first port is selected as the reference port and the spatial correlation between the k -th port and the reference port is characterized by Jakes' model. Based on this model, the authors derived the ergodic capacity and a corresponding lower

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62571430 and the Shaanxi Concept Demonstration Center under Grant No. 2025CG-GNYZ2-03.

bound. Subsequently, building directly upon the same reference point framework, the work in Ref. [7] derived the exact joint distribution of the signal envelopes for all ports, which enabled the derivation of the outage probability. Considering the outdated channel state information (CSI), the authors in Ref. [8] also employed the reference point model to develop an analytical framework for the outage probability of FAS. Although the reference point model has been widely used to derive expressions of the performance metric, it only captures the correlation of each port with a single reference and ignores correlations among other ports, leading to an overestimation of the achievable transmission performance. To address this limitation, Khammassi et al. introduced a fully correlated model^[9]. By performing an eigenvalue decomposition on the channel covariance matrix prescribed by Jakes' model, the fully correlated model in Ref. [9] expresses each channel coefficient as a weighted sum of the eigenvectors, where the weights are independent Gaussian random variables scaled by the square roots of the corresponding eigenvalues. This construction rigorously ensures that the correlation between any two ports is consistent with Jakes' model, thereby establishing a more accurate theoretical foundation for FAS performance analysis. Based on this model, the authors derived an approximate expression for the outage probability in single-integral form by retaining only the dominant eigenvalues of the channel covariance matrix to simplify the analysis^[9]. Although the fully correlated model addresses the limitations of the reference point model, it imposes high computational complexity. To address this issue, the authors in Ref. [10] proposed a spatial block-correlation model. The key idea of this model is the "structured approximation" using a block-diagonal matrix to capture the dominant spatial correlations. This construction is guided by the finding that the number of significant eigenvalues of the true covariance matrix is approximately $2W$ for a 1D fluid antenna of normalized length W . By grouping the ports into independent blocks, each designed to approximate one dominant eigenvalue, the model effectively decouples the system. This results in mutual independence between blocks and equal correlation within a block, thereby significantly simplifying subsequent mathematical analysis and computation. A persistent bottleneck across these models, however, is the lack of a closed-form expression for the outage probability. This is because the Rayleigh fading model leads to a probability density function (PDF) that involves an exponential term. When evaluating the outage probability for selection combining, this exponential term becomes coupled with the modified Bessel function I_0 , forming the integral kernel of the Marcum Q-function. It is the presence of this particular function within the integrals that prevents a closed-form evaluation. The lack of a closed-form solution not only hinders rapid performance evaluation but also becomes a bottleneck when extending FAS to more complex and promising frontiers. For instance, in multi-user systems, efficient power allocation and interference management rely on trac-

table performance metrics^[11]. Furthermore, the integration of FAS with advanced architectures, such as reconfigurable holographic surfaces, demands efficient analytical tools for joint optimization^[12]. Thus, the pursuit of a closed-form outage probability expression is motivated not only by theoretical elegance but also by practical necessity in advancing these cutting-edge applications.

To obtain a closed-form outage probability expression, we employ the same system model as in Ref. [10]. Within this framework, our main contributions are as follows:

- A novel approximation is introduced by applying an M -term Taylor series expansion to the zero-order modified Bessel function of the first kind, I_0 , which constitutes the core of the Marcum Q-function's integral kernel. The proposed closed-form approximation achieves high computational efficiency by eliminating numerical integration, enabling rapid parameter sweeps and real-time analysis. Furthermore, it provides tunable accuracy through adjustment of the M -term Taylor series, allowing for a flexible balance between precision and complexity.
- By integrating this approximation with the multinomial theorem, we derive an approximate closed-form expression for the outage probability, establishing an efficient and accurate tool for FAS performance evaluation.
- The theoretical findings are rigorously validated through Monte Carlo simulations, confirming the validity and precision of our analytical approach.

2 System Model

We consider the same point-to-point system model as in Ref. [10], with a single-user FAS receiver, as illustrated in Fig. 1.

The array consists of N uniformly spaced ports arranged linearly on a dielectric holder of length $W\lambda$, where λ denotes the wavelength of the transmitted signals.

We assume that all communication links experience Rayleigh fading. Consequently, the signal received at the k -th port of the fluid antenna can be expressed as

$$y_k = g_k x + \eta_k \quad (1),$$

where x is the transmitted data symbol with energy P , g_k is the

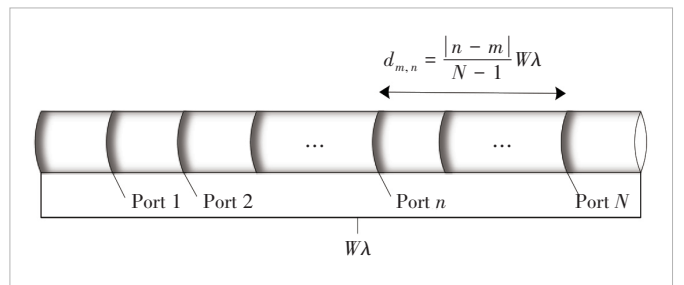


Figure 1. Geometry of the linear FAS with N candidate ports

channel coefficient at the k -th port with $E[|g_k|^2] = \sigma^2$, and $\eta_k \sim \mathcal{CN}(0, \sigma_\eta^2)$ is the additive white Gaussian noise (AWGN). Next, we define the average signal-to-noise ratio (SNR) as $\Gamma \triangleq \sigma^2 \frac{E[|x|^2]}{\sigma_\eta^2}$. For simplicity, and without loss of generality, we set $\sigma^2 = 1$ in the following derivations.

The spatial correlation among the FAS ports is characterized using the block-diagonal model of Ref. [10]¹, whose specific construction is detailed below.

1) Spatial coherence region partitioning

The total N FAS ports are partitioned into B independent spatial blocks, with each block containing $L(i)$ ports (satisfying $\sum_{i=1}^B L(i) = N$). The inter-port spacing within each block is smaller than the spatial coherence distance, while the spacing between blocks exceeds the coherence distance.

2) Intra-block correlation matrix

The covariance matrix of the i -th block follows a constant correlation structure:

$$\mathbf{A}_b = \begin{pmatrix} 1 & \mu_i^2 & \cdots & \mu_i^2 \\ \mu_i^2 & 1 & \cdots & \mu_i^2 \\ \vdots & \vdots & \ddots & \vdots \\ \mu_i^2 & \mu_i^2 & \mu_i^2 & 1 \end{pmatrix} \quad (2),$$

where $\mu_i^2 \rightarrow \Gamma$.

3) Inter-block independence

The global covariance matrix exhibits a block-diagonal structure, and this configuration models the FAS channel as B independent spatial degrees of freedom, where B is determined by the antenna aperture size W ($B \approx 2W$ for linear FAS)^[10].

3 Outage Performance Analysis

In this section, we derive a closed-form expression for the outage probability of FAS with optimal port selection. The system adopts selection combining (SC) that dynamically selects the antenna port exhibiting maximum instantaneous received signal power:

$$h_{\text{FAS}} = \max\{h_1, \dots, h_N\} \quad (3),$$

where $h_n = |g_n|^2$. Thus, we can write $\{h_n\}$ as

$$h_n = \left(\sqrt{1 - \mu_n^2} x_n + \mu_n x_0 \right)^2 + \left(\sqrt{1 - \mu_n^2} y_n + \mu_n y_0 \right)^2 \quad (4).$$

For the single-user scenario, the resulting outage probability is given by Eq. (43) in Ref. [10] as:

$$P_{\text{out}}(\gamma) = P\left(\max_n \{(1 - \mu^2)X_n\} < \gamma\right) \quad (5),$$

where X_n is defined as

$$X_n = \frac{h_n}{1 - \mu_n^2} = \left(x_n + \frac{\mu_n}{\sqrt{1 - \mu_n^2}} x_0 \right)^2 + \left(y_n + \frac{\mu_n}{\sqrt{1 - \mu_n^2}} y_0 \right)^2 \quad (6).$$

Based on the block-correlation framework established in Ref. [10], the outage probability of the considered FAS is given by Lemma 2 in Ref. [10] as:

$$P_{\text{out}}(\gamma) = \prod_{b=1}^B \int_0^\infty \frac{1}{2} e^{-\frac{r_0}{2}} \times \prod_{k=1}^{L(b)} \left[1 - Q_1 \left(\sqrt{\frac{\mu^2 r_0}{1 - \mu^2}}, \sqrt{\frac{\gamma}{1 - \mu^2}} \right) \right] dr_0 \quad (7).$$

However, the expression in Eq. (7) remains intractable for deriving a closed-form solution. The key obstacle is the presence of the Marcum Q-function $Q_1 \left(\sqrt{\frac{\mu^2 r_0}{1 - \mu^2}}, \sqrt{\frac{\gamma}{1 - \mu^2}} \right)$

within the integral, which itself is defined by an integral containing a Bessel function. To overcome this fundamental limitation, we propose a novel approximation method that yields a closed-form expression for the outage probability.

The primary challenge in obtaining a closed-form solution stems from the lack of standard integration techniques for the composite structure of the Marcum Q-function. Therefore, we expand it as follows:

$$P_{\text{out}}(\gamma) = \prod_{b=1}^B \int_0^\infty \frac{1}{2} e^{-\frac{r_0}{2}} \prod_{k=1}^{L(b)} \int_0^{\frac{\gamma}{1 - \mu^2}} \frac{1}{2} e^{-\frac{t_k + \frac{\mu^2}{1 - \mu^2} r_0}{2}} \times I_0 \left(\sqrt{\frac{\mu^2 r_0 t_k}{1 - \mu^2}} \right) dt_k dr_0 \quad (8).$$

The challenge of obtaining a closed-form expression persists with Eq. (8), again due to the non-elementary Bessel function I_0 and the nested integral. To overcome this, we apply a Taylor series expansion to $I_0(z)$. The general M -term Taylor series is given by:

$$I_0(z) = \sum_{m=0}^{M-1} \frac{z^{2m}}{2^{2m} (m!)^2} \quad (9).$$

¹ The block-diagonal model is chosen as it strikes a balance between accuracy and analytical tractability. While this model offers a representative structure for derivation, the proposed analytical framework, which employs Taylor series expansion and multinomial theorem to handle the outage probability integral, constitutes the primary contribution and is readily generalizable to other spatial correlation models.

By substituting this expansion into Eq. (8), we obtain a new expression that is amenable to analytical integration:

$$P_{\text{out}}(\gamma) \approx \prod_{b=1}^B \int_0^\infty \frac{1}{2} e^{-\frac{r_0}{2}} \prod_{k=1}^{L(b)} \int_0^{\frac{\gamma}{1-\mu^2}} \frac{1}{2} e^{-\frac{t_k + \frac{\mu^2}{1-\mu^2} r_0}{2}} \times \left[\sum_{m=0}^{M-1} \frac{1}{4^m (m!)^2} \left(\frac{\mu^2 r_0 t_k}{1-\mu^2} \right)^m \right] dt_k dr_0 \quad (10).$$

To illustrate the core idea with clarity and to derive a compact, computationally efficient approximation, we first present the specific case of retaining only the first two terms (i.e., $M = 2$) of the series:

$$P_{\text{out}}(\gamma) \approx \prod_{b=1}^B \int_0^\infty \frac{1}{2} e^{-\frac{r_0}{2}} \prod_{k=1}^{L(b)} \int_0^{\frac{\gamma}{1-\mu^2}} \frac{1}{2} e^{-\frac{t_k + \frac{\mu^2}{1-\mu^2} r_0}{2}} \times \left(1 + \frac{\mu^2 r_0 t_k}{4(1-\mu^2)} \right) dt_k dr_0 = \prod_{b=1}^B \int_0^\infty \frac{1}{2^{1+L(b)}} e^{-\left(\frac{r_0}{2} + \left(\frac{\mu^2 r_0}{2(1-\mu^2)} \right)^{L(b)} \right)} \prod_{k=1}^{L(b)} \left[\int_0^{\frac{\gamma}{1-\mu^2}} e^{-\frac{t_k}{2}} dt_k + \int_0^{\frac{\gamma}{1-\mu^2}} e^{-\frac{t_k}{2}} \frac{\mu^2 r_0 t_k}{4(1-\mu^2)} dt_k \right] dr_0 \quad (11).$$

Then we evaluate the inner integral:

$$\int_0^{x_k} e^{-\frac{t_k}{2}} dt_k = 2 \left(1 - e^{-\frac{x_k}{2}} \right) \int_0^{x_k} e^{-\frac{t_k}{2}} \frac{\mu^2 r_0 t_k}{4(1-\mu^2)} dt_k = \frac{\mu^2 r_0}{(1-\mu^2)} \left(1 - \left(1 + \frac{x_k}{2} \right) e^{-\frac{x_k}{2}} \right) \quad (12).$$

Substituting Eq. (12) into the outer integral:

$$P_{\text{out}}(\gamma) \approx \prod_{b=1}^B \frac{1}{2^{1+L(b)}} \int_0^\infty e^{-\left(\frac{1}{2} + \frac{\mu^2 L(b)}{2(1-\mu^2)} \right) r_0} \times \left[\prod_{k=1}^{L(b)} \left(2 \left(1 - e^{-\frac{\gamma}{2(1-\mu^2)}} \right) + \frac{\mu^2 r_0}{(1-\mu^2)} \left(1 - \left(1 + \frac{\gamma}{2(1-\mu^2)} \right) e^{-\frac{\gamma}{2(1-\mu^2)}} \right) \right) \right] dr_0 \quad (13).$$

For computational simplicity, we define $\alpha = \frac{1}{2} + \frac{\mu^2 L(i)}{2(1-\mu^2)}$, $D = \frac{\gamma}{2(1-\mu^2)}$, $E = 2(1 - e^{-D})$, $F = \frac{\mu^2}{(1-\mu^2)}(1 - (1 + D)e^{-D})$, so that

$$P_{\text{out}}(\gamma) = \prod_{b=1}^B \frac{1}{2^{1+L(b)}} \int_0^\infty e^{-\alpha r_b} [E + F r_b]^{L(b)} dr_b \quad (14).$$

Next, the complex integral is transformed into a manageable series through binomial expansion. After interchanging the summation and integration, the integral values are computed using the Gamma function. Based on the independence assumption, the final expression is obtained in a product form:

$$P_{\text{out}}(\gamma) \approx \prod_{b=1}^B \frac{1}{2^{1+L(b)}} \sum_{k=0}^{L(b)} \binom{L(b)}{k} E^{L(b)-k} F^k \frac{\Gamma(k+1)}{\alpha^{k+1}} \quad (15).$$

Having established the closed-form approximation for $M = 1$, we now extend the analysis to an arbitrary number of terms M . The derivation for the general case, starting from Eq. (10), is considerably more involved, as it requires additional analytical techniques, such as the application of the multinomial theorem, to handle the generalized series expansion. The resulting general closed-form expression for the outage probability is given by:

$$P_{\text{out}}(\gamma) = \prod_{b=1}^B \frac{1}{2^{L(b)+1}} \sum_{k_0+k_1+\dots+k_M=L(b)} \binom{L(b)}{k_0, k_1, \dots, k_M} \times \prod_{m=0}^{M-1} \frac{d_m^{k_m} \Gamma(s+1)}{\alpha^{s+1}} \quad (16),$$

where d_m represents the contribution weight corresponding to different expansion terms in the outage probability expression; s is a summation exponent, which computes the weighted sum of the orders of all Taylor expansion terms under a specific exponent combination $\{k_0, k_1, \dots, k_M\}$. The parameters in Eq. (16) are defined as follows:

$$d_m = \frac{2^{-m+1}}{(m!)^2} \left(\frac{\mu^2}{1-\mu^2} \right)^m \gamma(j+1, D), \quad s = \sum_{j=0}^{M-1} j k_j \quad (17).$$

The detailed proof of the above result is provided in Appendix A.

To quantitatively validate the low complexity of the proposed closed-form expression, we compare its computational efficiency with Eq. (7) via both theoretical analysis and MATLAB simulations. First, the computational complexity of Eq. (7) scales as $O\left(X_1 \cdot X_2 \cdot \sum_{b=1}^B L(b)\right)$, where X_1 and X_2 denote the numbers of discretization points for the outer and inner integrals (typically $X_1, X_2 \geq 10^3$ to ensure accuracy). In contrast, the complexity of Eq. (16) is given by $O\left(\binom{L(i)+M}{M} \cdot M\right)$. The binomial coefficient grows only polynomially with $L(i)$, leading to a substantially lower complexity compared with the numerical benchmark. Fig. 2 presents a runtime comparison between the proposed method (for N in

$\{5, 10\}$ and M in $\{2, 4, 6\}$ and the numerical benchmark in Eq. (7), conducted in MATLAB.

A direct comparison reveals that, for the same port count N , the runtime of the proposed closed-form method (Figs. 2a – 2c and 2e – 2g) is consistently and significantly lower than that of the numerical integration method (Figs. 2d and 2h), irrespective of the Taylor expansion term count M , strongly corroborating the theoretical complexity analysis.

To provide rigorous theoretical guarantees for the proposed approximation method, we present here the error bound analysis and convergence proof.

1) Error bound analysis

First, we define $T_M(x)$ as the Taylor expansion truncated after the M -th term:

$$T_M(x) = \sum_{m=0}^{M-1} \frac{x^{2m}}{2^{2m}(m!)^2} \quad (18).$$

The truncation error is defined as

$$\varepsilon_1(x, M) = I_0(x) - T_M(x) \quad (19).$$

Theorem 1 (Truncation Error Bound): For any $x \geq 0$ and positive integer M ,

$$\varepsilon_1(x, M) \leq \frac{x^{2M}}{2^{2M}(M!)^2} I_0(x) \quad (20).$$

2) Propagation to outage probability error

Let $P_{\text{out}}^{\text{exact}}$ denote the true outage probability under the block-correlation model, and $P_{\text{out}}^{(M)}$ denote the approximation using M Taylor terms.

Theorem 2 (Outage Probability Error Bound): Let $\varepsilon_{p_{\text{out}}}$ denote the outage probability error, which is defined as

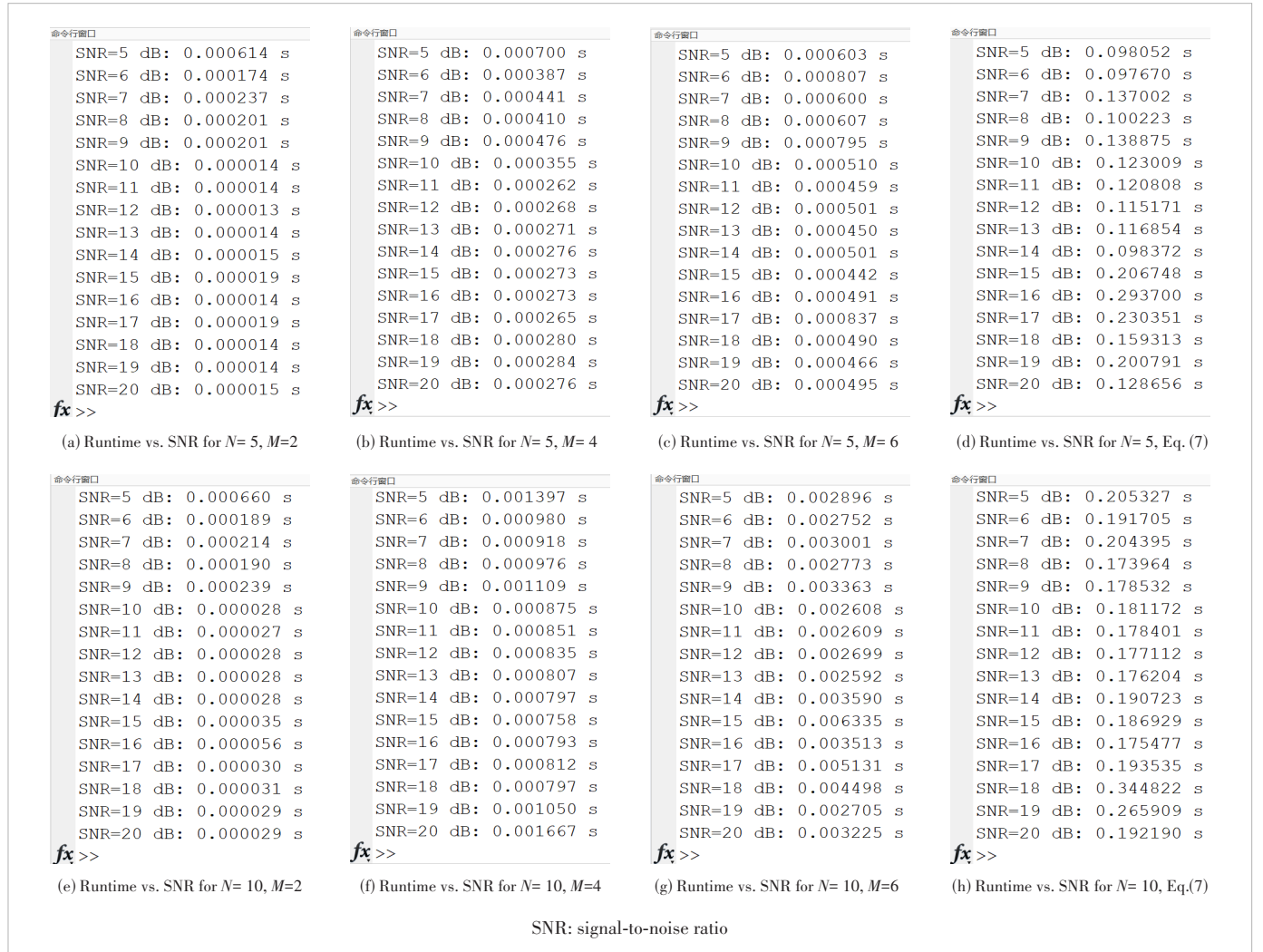


Figure 2. Runtime comparison for different numbers of ports and Taylor expansion terms versus SNR

$$\varepsilon_{P_{\text{out}}} = \left| P_{\text{out}}^{\text{exact}} - P_{\text{out}}^{(M)} \right| \leq \frac{\Gamma^M}{(M!)^2} \cdot C(M, \Gamma) \quad (21),$$

where

$$C(M, \Gamma) = \int_0^\infty e^{-2t} t^M I_0(2\sqrt{\Gamma t}) dt \quad (22),$$

is a finite constant for given system parameters.

3) Convergence proof

Theorem 3 (Convergence): As $M \rightarrow \infty$, the approximate outage probability converges to the true value:

$$\lim_{M \rightarrow \infty} P_{\text{out}}^{(M)} = P_{\text{out}}^{\text{exact}} \quad (23).$$

The detailed proof of the above theorem is provided in Appendix B.

4 Simulations

We now validate our theoretical analysis with computer simulations. For the sake of presentation, we consider the following parameters: $\mu^2 = 0.95$, $W = 1$, $\gamma = 2$, $N \in \{5, 10\}$.

Fig. 3 depicts the relative approximation error, which is defined as $\epsilon_{\text{rel}} = (\text{simulation result} - \text{analytical result}) / \text{simulation result}$ [13], between the logarithmic forms of the series expansion of I_0 and its original expression given in Eq. (7). This error is plotted against the number of terms M included in the expansion, for different SNR values.

The simulation results validate both the convergence and efficiency of the proposed method across different operating conditions. First, for all tested SNR conditions, the approximation error for both port configurations exhibits a consistent monotonic decrease as the number of Taylor series expansion terms increases, confirming the convergence of our approach. Moreover, a systematic downward shift in the error curves is observed as the SNR increases from 10 dB to 20 dB, indicating that better channel conditions allow the desired accuracy to be achieved with fewer expansion terms, thereby improving computational efficiency. In this study, we adopt a threshold of $\epsilon_{\text{rel}} = 10^{-3}$ for evaluating the accuracy of the proposed closed-form expression. This choice is grounded in rigorous engineering judgment: when the reference outage probability approaches the ultra-reliable communication target of 99.999%, this threshold ensures that the absolute approximation error does not exceed about 10^{-3} . Such an error margin is considerably smaller than typical engineering safety allowances and therefore does not compromise the core conclusion regarding whether the system meets the 99.999% reliability target. At the same time, this threshold strikes a balance between numerical rigor and practical utility. Fig. 4 illustrates how the relative error varies with SNR.

The simulation results confirm the effectiveness and conver-

gence of the proposed Taylor-series-based approximation. The approximation error decreases monotonically as the number of retained terms increases, and becomes negligible in the medium-to-high SNR regime.

Based on the above analysis, the system outage probability for different numbers of expansion terms ($M = 2, 4, 6$) is evaluated against the increasing SNR, as shown in Fig. 5.

The proposed approximations achieve remarkable accuracy with substantially reduced computational cost compared to the full model. The 4-term and 6-term expansions show excellent agreement with the reference model across the entire SNR range, outperforming the 2-term approximation, especially in the low-SNR regime. In medium to high SNR regions (approximately above 10 dB), both approximations consistently satisfy the ultra-reliable communication criterion of $P_{\text{out}} < 10^{-5}$ across the entire SNR range. Moreover, the consistent align-

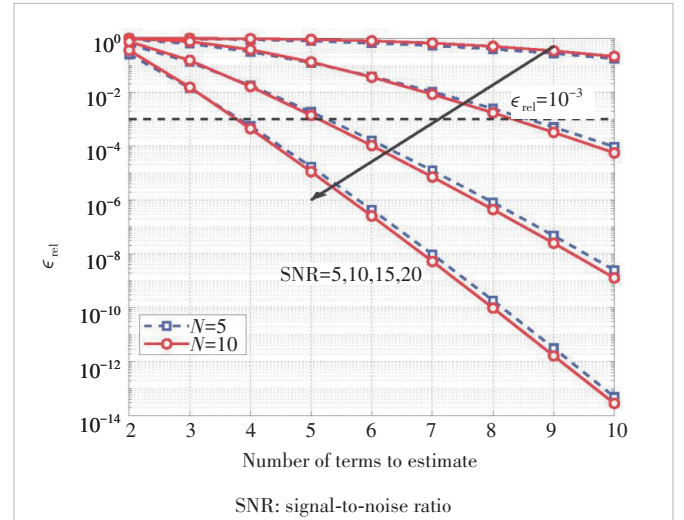


Figure 3. ϵ_{rel} versus the number of series expansion terms for I_0

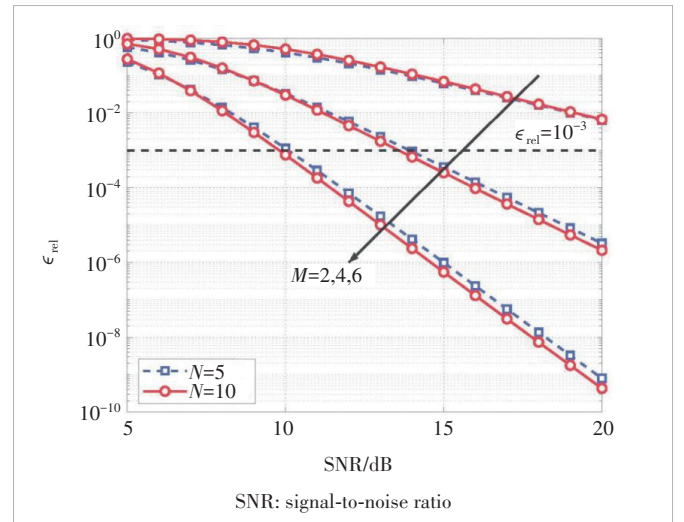


Figure 4. ϵ_{rel} versus SNR

ment across different port configurations ($N = 5$ and $N = 10$) further validates the adaptability and scalability of the proposed method.

5 Conclusions

In this paper, we tackle the challenging problem of deriving a closed-form expression for the outage probability of an FAS under the practical block-correlation model. The primary obstacle in this endeavor lies in the analytical intractability of the Marcum Q-function embedded within the outage probability integral. To overcome this challenge, our first key contribution is the introduction of a novel approximation for the Bessel function I_0 . By applying a precise M -term Taylor series expansion directly to its integral kernel, we successfully reformulated this function into a manageable M -term series representation. Building upon this mathematical foundation, our second major contribution was the successful derivation of an approximate closed-form expression for the FAS outage probability. By leveraging multinomial theorem and other analytical techniques, we consolidated the derived series into a comprehensive and computationally efficient formula.

Furthermore, the proposed approximation method for the Marcum Q-function is not limited to the block-correlation model and the Rayleigh channel. The core mathematical framework is generally applicable to both alternative spatial correlation models and broader classes of spatial fading models for FAS². Future work will focus on extending this analysis to multi-user FAS scenarios and incorporating more sophisticated fading channel models.

Appendix A

From (10), we can obtain the expression of $P_{\text{out}}(\gamma)$ like

$$P_{\text{out}}(\gamma) = \prod_{b=1}^B \int_0^\infty \frac{1}{2^{L(b)+1}} \exp\left(-r_0 \left(\frac{1}{2} + \frac{\mu^2 L(b)}{2(1-\mu^2)}\right)\right) \times \prod_{k=1}^{L(b)} \left[\sum_{m=0}^{M-1} \frac{1}{4^m (m!)^2} \left(\frac{\mu^2 r_0}{1-\mu^2}\right)^m \int_0^{\frac{\gamma}{1-\mu^2}} e^{-\frac{t_k}{2}} t_k^m dt_k \right] dr_0 \quad (24).$$

Then, using the lower incomplete Gamma function $\gamma(s, x) = \int_x^\infty t^{s-1} e^{-t} dt$, Eq. (24) becomes

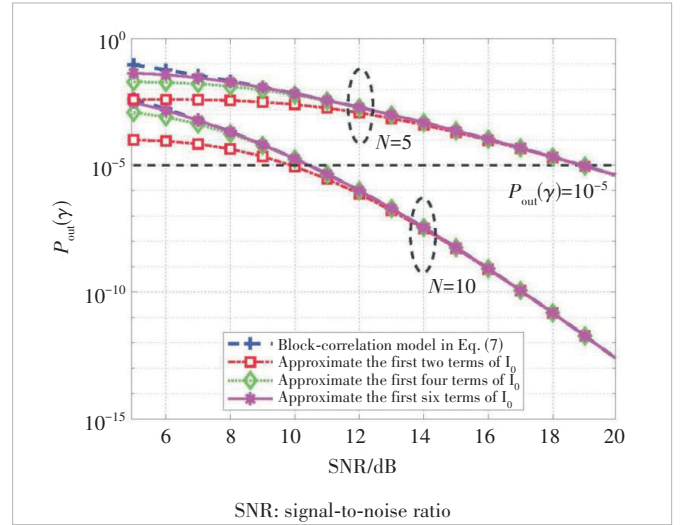


Figure 5. Outage probability versus SNR

$$P_{\text{out}}(\gamma) = \prod_{b=1}^B \int_0^\infty \frac{1}{2^{L(b)+1}} e^{-\alpha r_0} \left[\sum_{m=0}^{M-1} d_m r_0^m \right]^{L(b)} dr_0 \quad (25),$$

where d_m is given by Eq. (17).

To expand $\left[\sum_{m=0}^{M-1} d_m r_0^m \right]^{L(b)}$, we employ the multinomial expansion of a polynomial power, i. e. ,

$$\left[\sum_{m=0}^{M-1} d_m r_0^m \right]^{L(b)} = \sum_{k_0 + k_1 + \dots + k_M = L(b)} \binom{L(b)}{k_0, k_1, \dots, k_M} \times \prod_{j=0}^{M-1} d_j^{k_j} r_0^{\sum_{j=0}^{M-1} j k_j} \quad (26).$$

Then we can get

$$P_{\text{out}}(\gamma) = \prod_{b=1}^B \frac{1}{2^{L(b)+1}} \sum_{k_0 + k_1 + \dots + k_M = L(b)} \binom{L(b)}{k_0, k_1, \dots, k_M} \times \prod_{j=0}^{M-1} d_j^{k_j} \int_0^\infty e^{-\alpha r_0} r_0^s dr_0 \quad (27),$$

where s is given by Eq. (17). Using the Gamma function, we can get the final result like Eq. (16).

Appendix B

B.1 Proof of Theorem 1

Since the Taylor series of $I_0(x)$ is a positive series and its coefficient $a_k = 1/(k!)^2$ decays rapidly as k increases, it fol-

² This generality exists because the performance analysis of FAS under both alternative spatial correlation models (e.g., Jakes' model) and various fading models (e.g., Rician, Nakagami- m) consistently leads to probability expressions whose integral kernels involve special functions such as Bessel functions, Marcum Q-functions, or their variants^[14-15]. Precisely because the performance analysis under these different models ultimately yields integral expressions with similar kernels, the method proposed in our study can provide a unified processing strategy to transform such integrals into analytically tractable polynomial forms.

lows that:

$$\begin{aligned} \varepsilon_{I_0}(x, M) &= \sum_{k=M}^{\infty} \frac{x^{2k}}{2^{2k}(k!)^2} \leq \frac{x^{2M}}{2^{2M}(M!)^2} \sum_{k=0}^{\infty} \frac{x^{2k}}{2^{2k}(k!)^2} = \\ &= \frac{x^{2M}}{2^{2M}(M!)^2} I_0(x) \end{aligned} \quad (28).$$

B.2 Proof of Theorem 2

We substitute the error bound from Theorem 1 into the outage integral, interchange integral and summation, and bound the resulting expression using the exponential decay. Specific steps are as follows:

$$\begin{aligned} \varepsilon_{P_{out}} &= \left| \int_0^{\infty} e^{-2t} \varepsilon_{I_0}(2\sqrt{Ft}, M) dt \right| \leq \\ &= \int_0^{\infty} e^{-2t} \cdot \varepsilon_{I_0}(2\sqrt{Ft}, M) dt \leq \\ &= \int_0^{\infty} e^{-2t} \cdot \frac{(\sqrt{Ft})^{2M}}{(M!)^2} I_0(2\sqrt{Ft}) dt = \\ &= \frac{F^M}{(M!)^2} \int_0^{\infty} e^{-2t} t^M I_0(2\sqrt{Ft}) dt \end{aligned} \quad (29).$$

We define $C(M, F) = \int_0^{\infty} e^{-2t} t^M I_0(2\sqrt{Ft}) dt$ and then prove that it is bounded. The integrand is continuous for $t \geq 0$. As $t \rightarrow \infty$, we use the asymptotic expansion for I_0 :

$$I_0(2\sqrt{\rho t}) \sim \frac{e^{2\sqrt{\rho t}}}{\sqrt{4\pi\sqrt{\rho t}}} \quad (30).$$

So we obtain

$$e^{-2t} \cdot I_0(2\sqrt{Ft}) \sim \frac{e^{-2t+2\sqrt{Ft}}}{\sqrt{4\pi\sqrt{Ft}}} \quad (31).$$

The exponential term $-2t + \sqrt{2Ft}$ is dominated by $-2t$. For large t , the integrand decays exponentially, which results in overall exponential decay and ensures the convergence of the integral. Hence, $C(M, F)$ is bounded, completing the proof of Theorem 2.

B.3 Proof of Theorem 3

We define the following sequence of functions:

$$g_M(t) = e^{-2t} T_M(2\sqrt{Ft}) \quad (32).$$

Since I_0 is an entire function, its Taylor series converges uniformly on the entire complex plane, and therefore it also converges pointwise on the real axis. Hence, for any fixed $t \geq 0$ and $F \geq 0$, we have $\lim_{M \rightarrow \infty} T_M(2\sqrt{Ft}) = I_0(2\sqrt{Ft})$.

To apply the Lebesgue dominated convergence theorem, we construct a dominating function $G(t)$. Note that:

$$|g_M(t)| \leq e^{-2t} |T_M(2\sqrt{Ft})| \leq e^{-2t} I_0(2\sqrt{Ft}) = G(t) \quad (33).$$

We now prove that $G(t)$ is integrable on $[0, \infty)$. As $t \rightarrow \infty$,

$$G(t) = e^{-2t} \cdot I_0(2\sqrt{Ft}) \sim \frac{e^{-2t+2\sqrt{Ft}}}{\sqrt{4\pi\sqrt{Ft}}} \quad (34).$$

The exponential term $-2t + \sqrt{2Ft}$ is dominated by $-2t$ for large t , hence $G(t)$ decays exponentially, which guarantees the convergence of the integral. According to the Lebesgue dominated convergence theorem:

$$\lim_{M \rightarrow \infty} \int_0^{\infty} g_M(t) dt = \int_0^{\infty} \lim_{M \rightarrow \infty} g_M(t) dt = \int_0^{\infty} g(t) dt \quad (35).$$

So we obtain

$$\lim_{M \rightarrow \infty} P_{out}^{(M)} = 1 - \lim_{M \rightarrow \infty} \int_0^{\infty} g_M(t) dt = 1 - \int_0^{\infty} g(t) dt = P_{out}^{\text{exact}} \quad (36).$$

References

- [1] Popovski P, Nielsen J J, Stefanovic C, et al. Wireless access for ultra-reliable low-latency communication: principles and building blocks [J]. IEEE network, 2018, 32(2): 16 – 23. DOI: 10.1109/mnet.2018.1700258
- [2] Wang C X, You X H, Gao X Q, et al. On the road to 6G: visions, requirements, key technologies, and testbeds [J]. IEEE communications surveys & tutorials, 2023, 25(2): 905 – 974. DOI: 10.1109/comst.2023.3249835
- [3] He B, Xu H J, Ji W, et al. Reliable communication based on energy spreading transform and iterative detection in MIMO-OFDM systems [J]. IET communications, 2024, 18(14): 817 – 826. DOI: 10.1049/cmu2.12792
- [4] Heath R W, Lozano A. Foundations of MIMO communication [M]. Cambridge, UK: Cambridge University Press, 2018. DOI: 10.1017/9781139049276
- [5] Wong K K, Tong K F. Fluid antenna multiple access [J]. IEEE transactions on wireless communications, 2022, 21(7): 4801 – 4815. DOI: 10.1109/twc.2021.3133410
- [6] Wong K K, Shojaeifard A, Tong K F, et al. Performance limits of fluid antenna systems [J]. IEEE communications letters, 2020, 24(11): 2469 – 2472. DOI: 10.1109/lcomm.2020.3006554
- [7] Wong K K, Shojaeifard A, Tong K F, et al. Fluid antenna systems [J]. IEEE transactions on wireless communications, 2021, 20(3): 1950 – 1962. DOI: 10.1109/twc.2020.3037595
- [8] Psomas C, Kraidy G M, Wong K K, et al. Fluid antenna systems with outdated channel estimates [C]//Proceedings of ICC 2023 – IEEE International Conference on Communications. IEEE, 2023: 2970 – 2975. DOI: 10.1109/ICC45041.2023.10278619
- [9] Khammassi M, Kammoun A, Alouini M S. A new analytical approximation of the fluid antenna system channel [J]. IEEE transactions on wireless communications, 2023, 22(12): 8843 – 8858. DOI: 10.1109/TWC.2023.3266411
- [10] Ramírez-Espinosa P, Morales-Jimenez D, Wong K K. A new spatial block-correlation model for fluid antenna systems [J]. IEEE transactions on wireless communications, 2024, 23(11): 15829 – 15843. DOI: 10.1109/TWC.2024.3434509

- [11] Hu L Y, Li L C, Pan C H, et al. Two-timescale design for fluid antenna enhanced multiuser MIMO system with imperfect CSI [C]//Proceedings of IEEE 101st Vehicular Technology Conference (VTC2025-Spring). IEEE, 2025: 1 – 5. DOI: 10.1109/VTC2025-Spring65109.2025.11174765
- [12] Zhang S P, Zhang Y Z, Hashida H, et al. Fluid antenna systems enabled by reconfigurable holographic surfaces: beamforming design and experimental validation [J]. IEEE journal on selected areas in communications, 2026, 44: 1417 – 1431. DOI: 10.1109/JSAC.2025.3618797
- [13] Shi L Q, Ye Y H, Hu R Q, et al. System outage performance for three-step two-way energy harvesting DF relaying [J]. IEEE transactions on vehicular technology, 2019, 68(4): 3600 – 3612. DOI: 10.1109/TVT.2019.2897505
- [14] Huangfu J S, Song Z Y, Hou T W, et al. Performance analysis of fluid antenna system under spatially-correlated Rician fading channels [J]. IEEE transactions on wireless communications, 2026, 25: 1394 – 1407. DOI: 10.1109/twc.2025.3590722
- [15] Wang G Y, Zhang J, Xu K Z, et al. Performance analysis of multi-RIS-aided fluid antenna systems under Nakagami-m fading channels [C]//Proceedings of 16th International Conference on Wireless Communications and Signal Processing (WCSP). IEEE, 2024: 360 – 365. DOI: 10.1109/WCSP62071.2024.10826703

Biographies

Wu Ruojia is currently a master's candidate in the School of Communication

and Information Engineering, Xi'an University of Posts and Telecommunications, China. Her main research interests include wireless communications and fluid antenna systems.

Xu Rui received her BE and MS degrees from Xi'an Technological University, China and Xi'an University of Posts & Telecommunications, China in 2021 and 2024, respectively. She is currently pursuing a PhD degree at Xi'an University of Posts & Telecommunications. Her current research interests include backscatter communications and fluid antenna systems.

Ye Yinghui (connectyyh@126.com) holds a PhD and serves as an associate professor at the School of Communication and Information Engineering, Xi'an University of Posts & Telecommunications, China. His main research interests include simultaneous wireless information and power transfer (SWIPT), backscatter communication, and mobile edge computing.

Song Xi is currently a PhD candidate in the School of Electrical and Electronic Engineering, North China Electric Power University, China. His main research interests include backscatter communication, mobile edge computing, and simultaneous wireless information and power transfer (SWIPT).

Lu Guangyue holds a PhD and serves as a professor at the School of Communication and Information Engineering, Xi'an University of Posts & Telecommunications. His main research interests include communication signal processing, multi-antenna technology, and efficient spectrum utilization.

New Member of ZTE Communications Editorial Board



Zhang Wenjun

Zhang Wenjun is an Academician of the Chinese Academy of Engineering, an IEEE Fellow, a Chair Professor at Shanghai Jiao Tong University, China, and Director of the Cooperative Medianet Innovation Center (CMIC). He received the BS, MS, and PhD degrees in electronic engineering from Shanghai Jiao Tong University in 1984, 1987, and 1989, respectively. His research interests include image coding and communications, media-network convergence, broadband wireless transmission, and core technologies for digital television systems and chips. For more than three decades, he has led major national research and engineering programs and made systematic contributions to the development of indigenous technologies for television broadcasting and media networks in China. He has received four Second Prizes of the National Science and Technology Progress Award as the first-ranked contributor, the China Patent Gold Award, the Ho Leung Ho Lee Prize for Scientific and Technological Progress, and the Shanghai Science and Technology Meritorious Award.

Equivalent Degrees of Freedom Based Secrecy Outage Analysis for Fluid Antenna Wiretap Channels



Wu Tuo^{1,2}, Wu Zhenguan³, Zhi Kangda⁴, Liu Baiyang⁵,
Tong Kin-Fai^{2,5}

(1. South China University of Technology, Guangzhou 510640, China;
2. Shenzhen Research Institute, Hong Kong Metropolitan University,
Shenzhen 518057, China;
3. Dongguan University of Technology, Dongguan 523808, China;
4. Technical University of Berlin, Berlin 10587, Germany;
5. Hong Kong Metropolitan University, Hong Kong SAR, China)

DOI: 10.12142/ZTECOM.202603003

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260902.1035.002.html>,
published online September 2, 2026

Manuscript received: 2026-07-04

Abstract: Fluid antenna systems (FAS) improve wireless reliability by selecting the strongest port within a compact aperture, but the same port-selection capability also changes the physical-layer security behavior of a wiretap channel. Exact secrecy outage analysis is difficult because the legitimate and eavesdropping links each contain many spatially correlated ports, leading to high-dimensional joint distributions. This paper develops a closed-form secrecy outage framework for FAS wiretap channels using the equivalent degrees of freedom (EDoF) principle. Under the Jakes spatial correlation model, the many correlated physical ports are represented by a much smaller set of aperture-limited dominant spatial modes. The resulting model converts the correlated wiretap channel into a tractable competition between the effective selection branches available to Bob and Eve. An eigenvalue-weighted refinement further captures unequal dominant-mode powers. Closed-form and asymptotic results reveal how Bob's aperture determines secrecy diversity, how Eve's aperture strengthens the eavesdropping link, and why proportional Bob-Eve signal-to-noise ratio scaling produces a nonzero secrecy outage floor. Numerical simulations verify the analytical trends and show that the decisive security resource is Bob's aperture-induced EDoF advantage rather than the raw number of selectable ports.

Keywords: equivalent degrees of freedom; fluid antenna system; physical-layer security; secrecy outage probability; wiretap channel

Citation (Format 1): Wu T, Wu Z G, Zhi K D, et al. Equivalent degrees of freedom based secrecy outage analysis for fluid antenna wiretap channels [J]. *ZTE Communications*, 2026, 24(3): 12 – 17. DOI: 10.12142/ZTECOM.202603003

Citation (Format 2): T. Wu, Z. G. Wu, K. D. Zhi, et al., "Equivalent degrees of freedom based secrecy outage analysis for fluid antenna wiretap channels," *ZTE Communications*, vol. 24, no. 3, pp. 12 – 17, Sept. 2026. doi: 10.12142/ZTECOM.202603003.

1 Introduction

Physical layer security (PLS) exploits the randomness of wireless fading to protect confidential information directly at the physical layer^[1–2]. In a wiretap channel, secrecy is created by the instantaneous asymmetry between the legitimate receiver Bob and the eavesdropper Eve. This makes spatial diversity especially valuable, because a receiver that can sample more favorable fades than its adversary can reduce secrecy outage without necessarily increasing power or bandwidth.

Fluid antenna systems (FAS) provide a compact way to create such diversity^[3–8]. Recent hardware-oriented studies further support this practical motivation by showing that reconfigurable ports or activation states can be implemented within compact apertures, including millimeter-wave (mmWave) surface-wave-enabled prototypes, wideband pixel-based designs, multi-

activation structures with a single RF chain, and programmable meta-fluid antennas^[9–13]. An FAS exposes many candidate ports inside a finite aperture and activates the port with the most favorable channel realization^[3,14]. Compared with conventional multi-antenna arrays, this architecture can achieve strong selection gains with limited radio-frequency hardware, which makes it attractive for compact and space-constrained terminals^[8–13].

FAS-enabled secure transmission has already started to attract attention^[15–17], but rigorous secrecy analysis remains difficult. The central challenge is spatial correlation: the port gains of a FAS are strongly coupled through the aperture geometry and the underlying scattering field^[18–20]. Once Bob and Eve are both allowed to select their best ports, secrecy outage depends on the joint behavior of two maxima over correlated spatial samples, which quickly becomes analytically unwieldy.

This paper addresses that difficulty by transplanting the equivalent degrees of freedom (EDoF) viewpoint from outage

This work was partly funded by the Shenzhen Research Institute, Hong Kong Metropolitan University under Grant No. DON/2026/01 SZ.

analysis to the wiretap setting. The key idea is that the number of effective independent spatial modes is controlled by aperture size rather than by the nominal number of physical ports^[21–22]. Under the Jakes correlation kernel^[23–24], a one-dimensional aperture of normalized length W has approximately the following number of dominant modes:

$$K^* = 2\lceil W \rceil + 1 \quad (1)$$

Therefore, Bob's FAS can be interpreted as selection over K_B^* effective branches, while Eve's FAS can be interpreted as selection over K_E^* effective branches. This paper derives closed-form secrecy outage expressions, extends the method to a weighted independent modes (WIM) refinement, and turns the resulting formulas into aperture-based design insights.

2 FAS Wiretap System Model

2.1 Spatially Correlated FAS Channels

Consider a transmitter Alice sending confidential information to Bob in the presence of a passive Eve. Bob is equipped with a one-dimensional FAS containing N_B candidate ports over a normalized aperture W_B , and Eve with a one-dimensional FAS containing N_E candidate ports over a normalized aperture W_E .

The physical aperture lengths are L_B and L_E , and the normalized apertures are defined as $W_B = L_B/\lambda$ and $W_E = L_E/\lambda$, where λ is the carrier wavelength.

For each receiver q in $\{B, E\}$, the channel vector across all ports is modeled as follows.

$$\mathbf{g}_q = [g_{q,1}, \dots, g_{q,N_q}]^T \sim \text{CN}(0, \eta_q \mathbf{R}_q) \quad (2)$$

Equivalently, the generic model above specializes to the following complete Bob and Eve channel-vector constructions, where $\mathbf{z}_B$ and $\mathbf{z}_E$ are independent standard circularly symmetric complex Gaussian vectors:

$$\mathbf{g}_B = [g_{B,1}, \dots, g_{B,N_B}]^T = \sqrt{\eta_B} \mathbf{R}_B^{1/2} \mathbf{z}_B, \mathbf{z}_B \sim \text{CN}(0, \mathbf{I}_{N_B}),$$

$$\mathbf{g}_E = [g_{E,1}, \dots, g_{E,N_E}]^T = \sqrt{\eta_E} \mathbf{R}_E^{1/2} \mathbf{z}_E, \mathbf{z}_E \sim \text{CN}(0, \mathbf{I}_{N_E}).$$

Under the Jakes isotropic scattering model, the spatial correlation matrix is written as

$$[\mathbf{R}_q]_{m,n} = J_0 \left(\frac{2\pi W_q |m - n|}{N_q - 1} \right) \quad (3)$$

Bob and Eve each select their strongest available port, which leads to the following post-selection signal-to-noise ratios (SNRs).

$$\gamma_B = \bar{\gamma}_B X_B, \quad \gamma_E = \bar{\gamma}_E X_E \quad (4)$$

2.2 Secrecy Outage Probability

The instantaneous secrecy rate is defined by:

$$C_s = [\log_2(1 + \gamma_B) - \log_2(1 + \gamma_E)]^+ \quad (5)$$

For a target secrecy rate R_s , the secrecy outage probability is

$$P_{so} = P(C_s < R_s) \quad (6)$$

The outage event can be rewritten in the equivalent threshold form:

$$X_B < \frac{\theta(1 + \bar{\gamma}_E X_E) - 1}{\bar{\gamma}_B} \quad (7)$$

This threshold relation is the core object of the analysis because the secrecy outage is determined jointly by selected FAS gains of Bob and Eve.

For ease of reference, the symbol list is shown in Table 1.

3 EDoF Representation of FAS Wiretap Channels

3.1 KL Expansion and Spatial Mode Count

The channel correlation matrix admits an eigendecomposition, and the FAS channel can be represented through the Karhunen-Loève (KL) expansion as follows.

$$\mathbf{R}_q = \mathbf{U}_q \mathbf{\Lambda}_q \mathbf{U}_q^H, \mathbf{g}_q = \sqrt{\eta_q} \mathbf{U}_q \mathbf{\Lambda}_q^{1/2} \mathbf{z}_q \quad (8)$$

For the Jakes kernel, the dominant eigenvalues are concentrated in the first few modes, which yields the EDoF count:

$$K_q^* = 2\lceil W_q \rceil + 1 \quad (9)$$

For a finite set of physical ports, the dominant-mode count for both receivers is therefore given in closed form by $K_q^* =$

Table 1. Parameters and their denotation

Parameter	Denotation
$q \in \{B, E\}$	Receiver index (Bob/Eve)
N_q	Number of candidate ports
$W_q = L_q/\lambda$	Aperture length normalized by carrier wavelength
$\mathbf{g}_q$ and $\mathbf{R}_q$	Channel vector and spatial correlation matrix
η_q	Average channel power
$\bar{\gamma}_q$ and X_q	Average SNR and normalized selected gain
$\min\{N_q, 2\lceil W_q \rceil + 1\}$	Effective dominant-mode count
R_s and θ (2 raised to the power R_s)	Target secrecy rate and its threshold factor
P_{so}	Secrecy outage probability

SNR: signal-to-noise ratio

$$\min \{ N_q, 2\lceil W_q \rceil + 1 \}, q \in \{ B, E \}.$$

The outer minimum enforces the physical limit that the number of effective modes cannot exceed the number of sampled ports. In the dense-port regime considered in this paper, $N_q \geq 2\lceil W_q \rceil + 1$ and the expression reduces to $2\lceil W_q \rceil + 1$.

Increasing the port count beyond the number needed to resolve the aperture therefore does not create unlimited independent branches; it mainly oversamples the same spatial field.

3.2 EDoF Approximation

The EDoF model replaces the maximum over correlated port gains by the maximum over K_q^* independent unit-mean exponential gains.

$$X_q^{\text{EDoF}} = \max_{1 \leq k \leq K_q^*} Z_{q,k}, \quad Z_{q,k} \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(1) \quad (10).$$

Hence the cumulative distribution function and the probability density function become:

$$\begin{aligned} F_q(x) &= (1 - e^{-x})^{K_q^*}, \quad x \geq 0, \\ f_q(x) &= K_q^* e^{-x} (1 - e^{-x})^{K_q^* - 1} \end{aligned} \quad (11).$$

The main design implication is aperture-based: Bob's security advantage stems from a larger effective aperture rather than from raw port oversampling alone.

4 Closed-Form Secrecy Outage Under EDoF

Theorem 1. Under the EDoF model, the secrecy outage probability (SOP) can be written in the following closed form.

$$P_{\text{so}}^{\text{EDoF}} = \sum_{j=0}^{K_B^*} \binom{K_B^*}{j} (-1)^j \exp\left(-\frac{j(\theta-1)}{\bar{\gamma}_B}\right) \frac{(K_E^*)!}{\prod_{m=1}^{K_E^*} \left(m + \frac{j\theta\bar{\gamma}_E}{\bar{\gamma}_B}\right)} \quad (12).$$

The expression shows that Bob's effective mode count K_B^* enters through the outer selection expansion, while Eve's effective mode count K_E^* enters through the Laplace transform of the selected eavesdropping gain.

Corollary 1. If Eve has no FAS, then $K_E^* = 1$ and the SOP reduces to:

$$P_{\text{so}}^{\text{EDoF}} = \sum_{j=0}^{K_B^*} \binom{K_B^*}{j} \frac{(-1)^j \exp\left(-\frac{j(\theta-1)}{\bar{\gamma}_B}\right)}{1 + \frac{j\theta\bar{\gamma}_E}{\bar{\gamma}_B}} \quad (13).$$

5 Refined WIM Secrecy Outage

5.1 Eigenvalue-Weighted Effective Branches

The EDoF model treats the dominant modes as equal-power branches. A more accurate approximation is obtained by using

normalized dominant eigenvalue weights:

$$\beta_{q,k} = \frac{\lambda_{q,k} K_q^*}{N_q}, \quad X_q^{\text{WIM}} = \max_{1 \leq k \leq K_q^*} \beta_{q,k} Z_{q,k} \quad (14).$$

The corresponding WIM cumulative distribution function is

$$F_q^{\text{WIM}}(x) = \prod_{k=1}^{K_q^*} (1 - e^{-x/\beta_{q,k}}) \quad (15).$$

This refinement preserves the same aperture-based interpretation as the EDoF model while retaining the nonuniform powers of the dominant eigenmodes.

5.2 Quantitative Approximation Error

For the setting in Fig. 3, define the absolute approximation error as $\varepsilon = |P_{\text{so}}^{\text{model}} - P_{\text{so}}^{\text{MC}}|$. At -5 dB, the EDoF and WIM errors are 1.46×10^{-3} and 9.70×10^{-3} , respectively. Around the transition region at 1 dB, WIM reduces the error from 9.34×10^{-3} to 5.21×10^{-3} , a 44.2% reduction. At 9 dB, the corresponding errors are 1.42×10^{-3} and 4.40×10^{-3} . Thus, eigenvalue weighting improves the local transition-region fit but is not uniformly more accurate over all SNRs; the simpler EDoF model can be closer when the maximum-over-ports dependence dominates the eigenvalue imbalance. At 13 dB, the Monte Carlo (MC) SOP is 5.0×10^{-5} , while the EDoF and WIM absolute errors are 2.92×10^{-6} and 8.05×10^{-5} ; above 15 dB, the 80 000-trial MC curve reaches its resolution limit of 1.25×10^{-5} , so relative-error percentages are not statistically meaningful.

6 Asymptotic Secrecy Behavior and Design Insights

6.1 Fixed Eve SNR

Theorem 2. If Eve's average SNR is fixed and Bob's average SNR grows large, the secrecy outage behaves asymptotically as:

$$P_{\text{so}}^{\text{EDoF}} \sim \frac{G_{\text{sec}}}{\bar{\gamma}_B^{K_B^*}}, \quad G_{\text{sec}} = \mathbb{E}\left[\left(\theta(1 + \bar{\gamma}_E X_E) - 1\right)^{K_B^*}\right] \quad (16).$$

Therefore, the secrecy diversity order is K_B^* . Eve's aperture changes the coding-gain-like constant through the distribution of X_E , but it does not change the asymptotic slope when Eve's average SNR is fixed.

6.2 Proportional Bob and Eve SNRs

If Eve's average SNR scales proportionally with Bob's, then the secrecy outage converges to a nonzero floor:

$$\bar{\gamma}_E = \alpha \bar{\gamma}_B, \quad P_{\text{so},\infty}^{\text{EDoF}} = \sum_{j=0}^{K_B^*} \binom{K_B^*}{j} (-1)^j \frac{(K_E^*)!}{\prod_{m=1}^{K_E^*} (m + j\alpha\theta)} \quad (17).$$

This shows that the secrecy outage probability cannot be forced to zero by transmit-power growth alone when Bob and Eve scale together. The decisive design quantity is the EDoF pair (K_B^*, K_E^*) created by the two apertures.

The target secrecy rate shifts this floor through θ , where θ is 2 raised to the power of R_s . Since increasing R_s enlarges the outage event for every channel realization, the floor is monotonically nondecreasing in R_s . For example, with $(K_B^*, K_E^*) = (7, 3)$ and $\alpha=1$, the floor increases from 0.332 at $R_s=0.1$ bit/s/Hz to 0.467, 0.636, and 0.877 at $R_s=0.5$ bit/s/Hz, 1 bit/s/Hz, and 2 bit/s/Hz, respectively. Hence, a higher confidential-rate requirement must be supported by a larger Bob-side EDoF advantage or a lower relative Eve link quality.

The new insight of the proposed framework is that secrecy behaves in aperture steps rather than in raw-port steps. Once Bob crosses an integer aperture threshold, two additional effective modes appear immediately; once Eve crosses such a threshold, the wiretap link becomes measurably more dangerous.

7 Numerical Results

This section evaluates the closed-form formulas numerically and compares them with Monte Carlo simulations of the correlated FAS channels. The simulation setup adopts the same Jakes correlation model and the same secrecy-rate definition as those used in the analysis.

The figures in this paper are arranged from the most basic trend to the most adversarial setting. They first show the effect of Bob's average SNR and aperture, and then illustrate the Bob-Eve aperture competition. Next, they illustrate the agreement among EDoF, WIM, and exact correlated-channel simulation, and finally demonstrate the high-SNR secrecy floor.

Compared with existing secrecy analyses that rely mainly on high-dimensional simulation, fitted correlation parameters, or nominal port count, the proposed EDoF/WIM framework expresses secrecy directly in terms of the aperture-induced mode counts and dominant eigenvalue weights of Bob and Eve. This yields clearer design rules and lower evaluation complexity.

7.1 Impact of Bob's Aperture

Fig. 1 verifies the secrecy diversity statement by varying Bob's average SNR under several Bob apertures. Increasing W_B enlarges K_B^* and steepens the outage decay. The figure shows that secrecy gain is achieved in aperture-controlled steps rather than by merely adding more densely packed ports.

7.2 Bob-Eve Aperture Competition

Fig. 2 sweeps Bob's normalized aperture W_B parameterized by Eve's aperture W_E . The staircase follows from $K_B^* = \min\{N_B, 2\lceil W_B \rceil + 1\}$: within one integer aperture interval, denser sampling resolves the same spatial field and does not create another independent mode. Crossing the next wavelength-normalized boundary expands the resolvable spatial bandwidth and introduces a conjugate pair of spatial

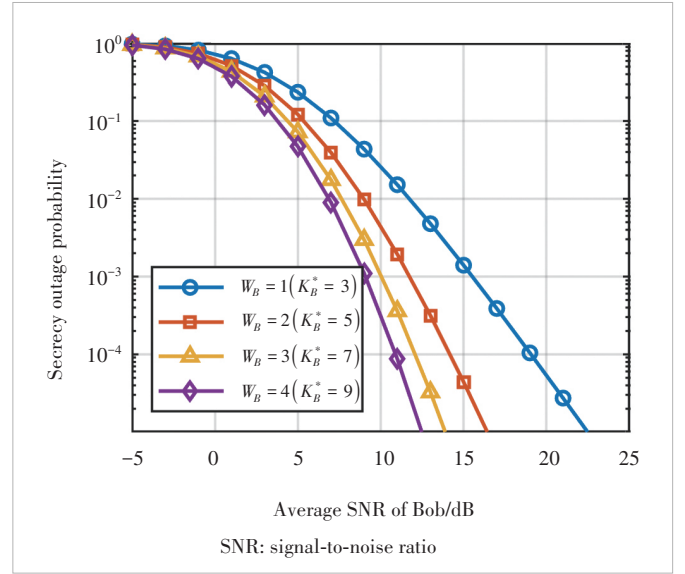


Figure 1. Secrecy outage probability versus Bob's average SNR for $W_E = 1$, $\gamma_E = 0$ dB, and $R_s = 0.5$ bit/s/Hz

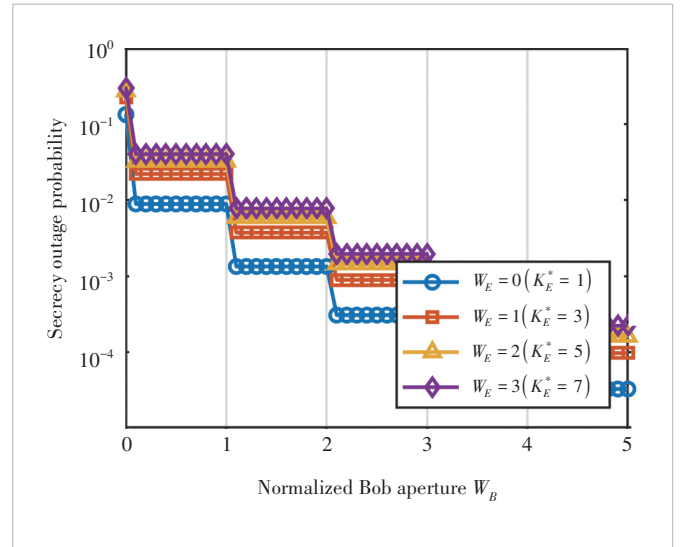


Figure 2. Secrecy outage probability versus Bob's normalized aperture for $\gamma_B = 15$ dB, $\gamma_E = 5$ dB, and different Eve apertures

modes, such that K_B^* increases by 2 and the SOP drops abruptly. The curves therefore expose the Bob-Eve aperture gap more directly than a raw port-count description.

7.3 EDoF and WIM Validation

Fig. 3 compares the EDoF and refined-WIM formulas with exact Monte Carlo simulation results for correlated FAS channels. Both analytical curves capture the SNR trend. The WIM model explicitly accounts for nonuniform dominant eigenvalues and improves the fit near the 1 dB transition point, whereas the equal-power EDoF approximation is closer at some other SNRs. This is consistent with the quantitative error audit in Section 5.2.

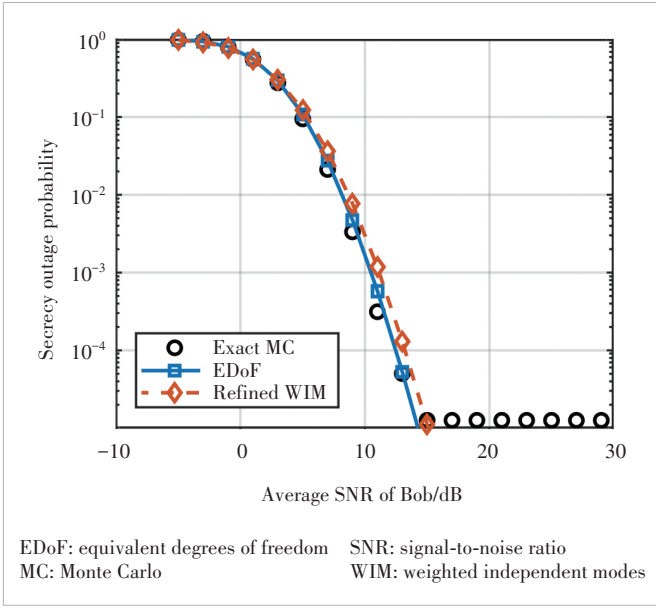


Figure 3. EDoF and refined-WIM secrecy outage validation against exact Monte Carlo simulation for $N_B = 40$, $N_E = 30$, $W_B = 3$, $W_E = 2$, and $\bar{\gamma}_E = 0$ dB

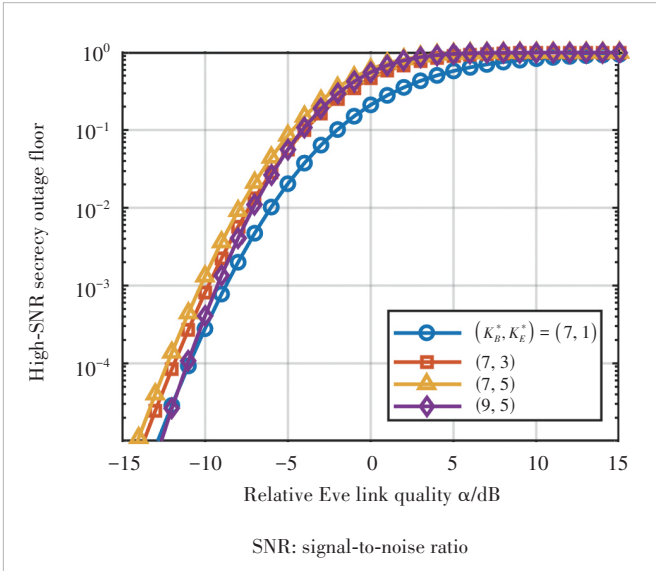


Figure 4. High-SNR secrecy outage floor versus the relative Eve link quality α for different EDoF pairs (K_B^*, K_E^*) and $R_s = 0.5$ bit/s/Hz

7.4 High-SNR Secrecy Floor

Fig. 4 visualizes the proportional-SNR secrecy floor versus the relative Eve link quality α . The floor rises with K_E^* and falls with K_B^* . At $\alpha = 0$ dB and $R_s = 0.5$ bit/s/Hz, increasing the EDoF pair from $(7, 5)$ to $(9, 5)$ lowers the floor from 0.611 to 0.558, while reducing Eve's mode count from five to three at $K_B^* = 7$ lowers it further to 0.467. Therefore, under a benchmark constraint $W_B + W_E = W_T$, the floor is minimized by assigning the controllable aperture budget to Bob until the next integer W_B

threshold is crossed, rather than spending it on additional ports inside the same interval. In practice, Eve's aperture is not controlled by the legitimate designer; W_E should be treated as a threat upper bound, and the actionable rule is to enlarge Bob's aperture (or constrain the region in which Eve can deploy) until K_B^* exceeds the anticipated K_E^* by the largest feasible margin.

8 Conclusions

This paper presents a standalone EDoF-based secrecy outage framework for FAS wiretap channels. By using the spectral concentration of the Jakes correlation kernel, a correlated FAS with many physical ports is reduced to a selection system with $K_q^* = \min\{N_q, 2\lceil W_q \rceil + 1\}$ effective branches. The results show that Bob's aperture controls secrecy diversity, Eve's aperture controls the strength of the selected wiretap link, and proportional Bob-Eve SNR scaling creates a nonzero secrecy outage floor. The Jakes model assumes a one-dimensional aperture, isotropic two-dimensional scattering, identical diffuse angular statistics, and ideal narrowband stationarity; its compact EDoF law therefore need not remain exact in directional, clustered, nonstationary, blocked, or three-dimensional urban channels. Future work will replace the Jakes kernel with correlation matrices obtained from standardized 3GPP urban micro-cell and urban macrocell channel models, extract effective mode counts from their eigenvalue spectra, and test whether the same EDoF/WIM secrecy structure extends to wideband, planar-aperture, mutual-coupling, and hardware-impaired FAS implementations. Overall, the central security resource is the aperture-induced EDoF advantage rather than the raw number of selectable ports.

References

- [1] Wyner A D. The wire-tap channel [J]. The Bell system technical journal, 1975, 54(8): 1355 – 1387. DOI: 10.1002/j.1538-7305.1975.tb02040.x
- [2] Bloch M, Barros J. Physical-layer security: from information theory to security engineering [M]. Cambridge, UK: Cambridge University Press, 2011. DOI: 10.1017/cbo9780511977985
- [3] Wong K K, Shojaeifard A, Tong K F, et al. Fluid antenna systems [J]. IEEE transactions on wireless communications, 2021, 20(3): 1950 – 1962. DOI: 10.1109/TWC.2020.3037595
- [4] Wong K K, Shojaeifard A, Tong K F, et al. Performance limits of fluid antenna systems [J]. IEEE communications letters, 2020, 24(11): 2469 – 2472. DOI: 10.1109/LCOMM.2020.3006554
- [5] New W K, Wong K K, Xu H, et al. A tutorial on fluid antenna system for 6G networks: encompassing communication theory, optimization methods and hardware designs [J]. IEEE communications surveys & tutorials, 2025, 27(4): 2325 – 2377. DOI: 10.1109/COMST.2024.3498855
- [6] New W K, Wong K K, Wang C, et al. Fluid antenna systems: redefining reconfigurable wireless communications [J]. IEEE journal on selected areas in communications, 2026, 44: 1013 – 1044. DOI: 10.1109/JSAC.2025.3632097
- [7] Hong H J, Wong K K, Chae C B, et al. A contemporary survey on fluid an-

- tenna systems: fundamentals and networking perspectives [J]. *IEEE transactions on network science and engineering*, 2026, 13: 2305 – 2328. DOI: 10.1109/TNSE.2025.3613225
- [8] Wu T, Zhi K D, Yao J T, et al. Fluid antenna systems enabling 6G: principles, applications, and research directions [J]. *IEEE wireless communications*, 2026, 33(4): 100 – 108. DOI: 10.1109/MWC.2025.3629597
- [9] Tong K F, Liu B Y, Wong K K. Designs and challenges in fluid antenna system hardware [J]. *Electronics*, 2025, 14(7): 1458. DOI: 10.3390/electronics14071458
- [10] Shen Y J, Tang B Y, Gao S, et al. Design and experimental validation of mmWave surface-wave-enabled fluid antennas for future wireless communications [J]. *IEEE antennas and wireless propagation letters*, 2026, 25(4): 1467 – 1471. DOI: 10.1109/LAWP.2026.3657059
- [11] Liu B Y, Wu T, Wong K K, et al. Wideband pixel-based fluid antenna system: an antenna design for smart city [J]. *IEEE Internet of Things journal*, 2026, 13(4): 6850 – 6862. DOI: 10.1109/IIOT.2025.3637319
- [12] Huang J W, Meng H, Liu B Y, et al. Multi-activation fluid antenna system with a single RF chain for fast multi-user access [C]//*Proceedings of International Symposium on Antennas and Propagation (ISAP)*. IEEE, 2025: 1 – 2. DOI: 10.23919/ISAP63122.2025.11362065
- [13] Liu B Y, Tong K F, Wong K K, et al. Programmable meta-fluid antenna for spatial multiplexing in fast fluctuating radio channels [J]. *Optics express*, 2025, 33(13): 28898. DOI: 10.1364/oe.567923
- [14] Chai Z, Wong K K, Tong K F, et al. Port selection for fluid antenna systems [J]. *IEEE communications letters*, 2022, 26(5): 1180 – 1184. DOI: 10.1109/LCOMM.2022.3152451
- [15] Tang B Y, Xu H, Wong K K, et al. Fluid antenna enabling secret communications [J]. *IEEE communications letters*, 2023, 27(6): 1491 – 1495. DOI: 10.1109/LCOMM.2023.3264634
- [16] Rostami Ghadi F, Wong K K, López-Martínez F J, et al. Physical layer security over fluid antenna systems: secrecy performance analysis [J]. *IEEE transactions on wireless communications*, 2024, 23(12): 18201 – 18213. DOI: 10.1109/TWC.2024.3463488
- [17] New W K, Wong K K, Xu H, et al. Fluid antenna system: new insights on outage probability and diversity gain [J]. *IEEE transactions on wireless communications*, 2024, 23(1): 128 – 140. DOI: 10.1109/TWC.2023.3276245
- [18] Wong K K, Tong K F, Chen Y, et al. Closed-form expressions for spatial correlation parameters for performance analysis of fluid antenna systems [J]. *Electronics letters*, 2022, 58(11): 454 – 457. DOI: 10.1049/el2.12487
- [19] Khammassi M, Kammoun A, Alouini M S. A new analytical approximation of the fluid antenna system channel [J]. *IEEE transactions on wireless communications*, 2023, 22(12): 8843 – 8858. DOI: 10.1109/TWC.2023.3266411
- [20] Vega-Sánchez J D, Urquiza-Aguilar L F, Mora H R C, et al. Fluid antenna system: secrecy outage probability analysis [J]. *IEEE transactions on vehicular technology*, 2024, 73(8): 11458 – 11469. DOI: 10.1109/TVT.2024.3376475
- [21] Slepian D, Pollak H O. Prolate spheroidal wave functions, Fourier analysis and uncertainty: I [J]. *The bell system technical journal*, 1961, 40(1): 43 – 63. DOI: 10.1002/j.1538-7305.1961.tb03976.x
- [22] Landau H J, Pollak H O. Prolate spheroidal wave functions, Fourier analysis and uncertainty: II [J]. *The bell system technical journal*, 1961, 40(1): 65 – 84. DOI: 10.1002/j.1538-7305.1961.tb03977.x
- [23] Jakes W C. *Microwave mobile communications* [M]. New York: Wiley, 1974
- [24] Wu T, Luk K M, Tang J, et al. Variable block-correlation modeling and optimization for secrecy analysis in fluid antenna systems [PP/OL]. *arXiv*. (2025-10-04) [2026-06-06]. <https://arxiv.org/abs/2510.03594>

Biographies

Wu Tuo received his BE degree in telecommunication engineering from South China Normal University, in 2017, MS degree in radio physics from Sun Yat-sen University, China in 2021, and PhD degree from the School of Electronic Engineering and Computer Science, Queen Mary University of London, UK in 2024. He was a postdoctoral researcher with Nanyang Technological University, Singapore from 2024 to 2025, and a research fellow with City University of Hong Kong, China from 2025 to 2026. He is currently an associate professor with the School of Electronic and Information Engineering, South China University of Technology. His research interests include fluid antenna systems, reconfigurable intelligent surfaces, wireless localization, and sensing. He has published extensively in wireless communications and serves on the editorial boards of several IEEE journals.

Wu Zhenguan received his BE degree in communication engineering from Zhaoqing University, China in 2025. He is currently a graduate student at Dongguan University of Technology, China. His research interests include fluid antenna systems and reconfigurable intelligent surfaces.

Zhi Kangda received his BE degree from Shanghai University, China in 2017, ME degree from the University of Science and Technology of China in 2020, and PhD degree from Queen Mary University of London, UK in 2023. He is currently a Marie Curie Research Fellow with the Technical University of Berlin, Germany. His research interests include next-generation MIMO systems, reconfigurable intelligent surfaces, sensing and estimation, and near-field communications. He received the 2026 IEEE ComSoc CTTC Andrea Goldsmith Young Scholars Award, the 2025 IEEE ComSoc Heinrich Hertz Award for Best Communications Letter, the 2025 IEEE ComSoc EMEA Outstanding Young Researcher Award, and the 2024 Marie Curie Postdoctoral Fellowship. He serves as an editor for several IEEE journals and as a Guest Editor for *IEEE Network* and *EURASIP Journal on Wireless Communications and Networking*.

Liu Baiyang received his BS degree in wireless communication from Guangdong University of Technology, China, and his PhD degree in communication and information systems from South China University of Technology in 2018. He is currently an assistant professor with the College of Artificial Intelligence, Shenzhen Technology University, China. He has held research fellow positions at Hong Kong Metropolitan University, China; City University of Hong Kong, China; Queen Mary University of London, UK; Southern University of Science and Technology, China; and Shenzhen University, China. His research interests include reconfigurable intelligent surfaces, fluid antenna systems, orbital angular momentum multiplexing, and integrated sensing and communication. He has published extensively in antennas and wireless communications, and received the Best Paper Award at GSMW 2024 and the 2024 Best Reviewer Awards from *IEEE Transactions on Antennas and Propagation* and *IEEE Antennas and Wireless Propagation Letters*.

Tong Kin-Fai (ktong@hkmu.edu.hk) is currently a Chair Professor of antennas and applied electromagnetics at Hong Kong Metropolitan University, China. His research interests include wideband antennas, circularly polarized patch antennas, microwave and millimeter-wave techniques, fluid antennas, surface waves, and wideband mmWave transducers for future wireless communications systems. He has made significant contributions to research translation through industrial collaboration, patents, commercialization, and public engagement. His Innovate UK project was rated “outstanding” and featured in the *Parliamentary Review*. He received the UCL Knowledge Transfer Business of the Year Award in 2015 and the UCL Provost’s Spirit of Enterprise Award in 2017. He is a Chartered Engineer (UK), a Fellow of the Electromagnetics Academy, and serves as an editor for several leading journals in antennas, electromagnetics, and wireless propagation.

Prediction-Window Aided Fluid Antenna Systems for High-Mobility Communications



Huang Yinning¹, Zeng Cheng¹, Yang Zhaohui²

(1. Nanjing University of Science and Technology, Nanjing 210094, China;
2. Zhejiang University, Hangzhou 310058, China)

DOI: 10.12142/ZTECOM.202603004

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260910.1053.002.html>,
published online September 10, 2026

Manuscript received: 2026-06-17

Abstract: Fluid antenna systems (FAS) provide continuous spatial degrees of freedom for next-generation wireless networks. However, deploying FAS on high-speed mobile embodied intelligent agents introduces Doppler shifts that shorten the channel coherence time, causing conventional full-port scanning to incur excessive pilot overhead. To address this issue, this paper proposes a prediction-aided adaptive communication architecture that constructs a dynamic local sounding window centered on upper-layer prediction results. By exploiting the spatial channel correlation among FAS ports, the proposed approach balances pilot overhead and spatial diversity gain under prediction deviations. We establish a quantitative model for the long-term expected throughput by integrating a nonlinear bounded hit probability function and a gain bridging model based on the zeroth-order Bessel function of the first kind. We then formulate a discrete optimization problem under practical constraints to maximize the expected transmission capacity, and develop a two-stage low-complexity algorithm that decouples the online computational complexity from the number of antenna ports. Simulation results show that the proposed mechanism adaptively adjusts the sounding window size according to moving velocity and prediction accuracy, reducing pilot overhead in high-mobility scenarios and achieving higher throughput than fixed-window and global scanning schemes.

Keywords: fluid antenna system; embodied intelligent agent; overhead catastrophe; prediction window; discrete zeroth-order gradient optimization

Citation (Format 1): Huang Y N, Zeng C, Yang Z H. Prediction-window aided fluid antenna systems for high-mobility communications [J]. ZTE Communications, 2026, 24(3): 18 – 30. DOI: 10.12142/ZTECOM.202603004

Citation (Format 2): Y. N. Huang, C. Zeng, and Z. H. Yang, "Prediction-window aided fluid antenna systems for high-mobility communications," *ZTE Communications*, vol. 24, no. 3, pp. 18 – 30, Sept. 2026. doi: 10.12142/ZTECOM.202603004.

1 Introduction

With the evolution of 6G networks toward space-air-ground-sea integration and autonomous operation, embodied intelligent agents with autonomous sensing, decision-making, and execution capabilities are gradually becoming core nodes at the network edge^[1–2]. In typical applications of vertical industries such as industrial inspection and environmental monitoring, such mobile agents need to complete continuous tasks in highly dynamic physical environments, which imposes stringent requirements on the reliability and effective throughput of wireless communication links.

Fluid antenna systems (FAS), as a revolutionary reconfigurable antenna technology, enable the radiating element to slide freely within a fixed physical aperture, thereby obtaining continuous spatial degrees of freedom in a very compact size to ac-

tively adapt to time-varying wireless channel states^[3–5]. Compared with conventional fixed antenna arrays, FAS can achieve superior channel gain in complex multipath environments due to its flexible spatial reconfiguration capability, and has demonstrated remarkable performance advantages in scenarios such as multi-user access interference suppression and integrated sensing and communication (ISAC). It has thus become a key research direction for next-generation wireless networks^[6–8].

However, existing FAS-related studies are generally based on the underlying assumption that user nodes are stationary or quasi-static, which presupposes that the channel coherence time is sufficient to support the full-port pilot sounding process. In the field of channel estimation and port selection, existing solutions include linear minimum mean square error (LMMSE), compressed sensing, and sparse Bayesian learning^[9–11], as well as various deep learning-assisted schemes^[12–14], all of which are validated under quasi-static block fading channel scenarios. When FAS is deployed on a high-speed moving embodied intelligent agent, the mechanical movement of the agent will induce severe Doppler shift, causing the channel coherence time to shrink sharply and non-linearly as the moving speed increases. If the global blind scanning mechanism that traverses all virtual

This work was partially supported by the National Natural Science Foundation of China under Grant No. 62601386, the Basic Research Program of Jiangsu Province under Grant No. BK20251413, the State Key Laboratory of Millimeter Waves under Grant No. K202602-01, and the Fundamental Research Funds for the Central Universities under Grant No. 30925010508.

ports is still adopted, the pilot sounding overhead will exhaust all coherence time slots, leaving zero time for effective data transmission and resulting in the performance dilemma of “overhead catastrophe”^[15].

To address this challenge, the most straightforward solution in existing research is to compress sounding overhead based on port prediction: By mining the temporal evolution patterns of channels or environmental sensing information, the optimal port position at the next moment is directly predicted, reducing the original N full sounding operations to single-point sounding, which fundamentally cuts down the pilot temporal overhead^[16]. Traditional methods include the Vec Prony algorithm and the matrix pencil moving port (MPMP) algorithm, which realize port tracking through angle-delay domain features. With the expanding application of large language models (LLMs) in physical layer communications, the state-of-the-art Port-LLM model, built on the GPT-2 backbone with a Low-Rank Adaptation (LoRA) fine-tuning architecture, achieves prediction of future optimal ports through temporal extrapolation of historical channel tables, and outperforms traditional prediction algorithms in spectral efficiency under medium- and high-speed scenarios^[17]. All the above studies have verified the significant overhead compression effect of port prediction, which serves as the core technical foundation for solving the overhead catastrophe in highly dynamic scenarios.

Nevertheless, existing port prediction studies generally have two limitations. First, they aim to output a single optimal port, assuming the prediction result is completely accurate without considering the performance degradation caused by prediction deviation. In fact, any prediction mechanism has errors, and pure single-point prediction has zero fault tolerance capability. Once the predicted port deviates from the true optimal port, the communication gain will directly drop to the channel level at the deviated position, with no buffer margin^[18]. Second, existing studies have not established a joint tradeoff mechanism among prediction confidence, moving speed, sounding overhead, and spatial diversity gain: They either adopt extremely compressed single-point sounding at the cost of robustness, or fall back to full scanning at the cost of transmission efficiency, failing to adaptively adjust the balance between overhead and fault tolerance according to real-time dynamic scenarios.

From the perspective of physical essence, there exists a gain spillover effect caused by spatial correlation among FAS ports^[19]: When there is a small deviation between the true optimal port and the sounding port, the channel gain does not drop sharply to zero, but presents a smooth attenuation characteristic constrained by the zeroth-order Bessel function of the first kind over spatial distance. This physical property implies that moderately expanding the sounding range on the basis of single-point prediction can cover high-probability prediction deviation intervals and obtain significant fault tolerance gain with only a tiny amount of additional pilot overhead. Conversely, completely abandoning window expansion and relying solely on single-point

prediction is equivalent to actively wasting the inherent spatial fault tolerance potential of FAS; yet blindly expanding the window will offset the overhead reduction benefit brought by prediction and slide back to the overhead catastrophe. Therefore, a key issue for further enhancing the performance robustness of highly dynamic FAS systems is how to design an adaptive local sounding mechanism based on the overhead compression capability of port prediction, while improving prediction fault tolerance with controllable additional overhead.

To fill the above research gap, this paper proposes a prediction-aided fault-tolerant window-enhanced adaptive communication mechanism for fluid antenna systems. The proposed mechanism assumes the existence of a generic port prediction module at the upper layer, whose prediction accuracy is uniformly characterized by a baseline hit rate parameter. Instead of focusing on the specific implementation of the prediction algorithm, this paper concentrates on the adaptation enhancement layer between the prediction results and the underlying hardware. Specifically, a local search window containing K consecutive virtual ports is constructed around the predicted anchor port. By leveraging the inherent spatial channel correlation of FAS to bridge prediction deviations, this design achieves substantial fault tolerance and spatial diversity gains at the expense of minimal and controllable pilot overhead. Furthermore, we designate the window size K as the core control variable to conduct a joint optimization coupled with the terminal moving velocity and the baseline prediction accuracy. In extreme high-speed and high-Doppler scenarios, the algorithm adaptively shrinks the window to $K = 1$ to ensure effective data transmission duration under critically short coherence times. In medium- and low-speed scenarios, the window is moderately expanded to maximize fault-tolerance robustness and channel capacity while keeping the pilot overhead affordable. Consequently, an optimal dynamic balance between pilot overhead and spatial diversity gain is achieved, fundamentally mitigating the overhead catastrophe in highly dynamic environments.

The main contributions of this paper are summarized as follows:

1) We propose a fundamentally new overhead management paradigm for FAS in highly dynamic scenarios. By constructing a dynamic local sounding window anchored on upper-layer port prediction results, we transform the conventional global blind search into a spatially localized precision sounding process. This architecture effectively compresses the pilot temporal overhead while robustly retaining critical fault-tolerance capabilities against algorithmic prediction deviations.

2) We establish a comprehensive mathematical model to evaluate the long-term expected throughput of the proposed system. To quantitatively characterize the intrinsic relationship between prediction accuracy and search window size, we introduce a nonlinear bounded-growth hit probability function. Furthermore, we derive an equivalent gain bridging model based on the zeroth-order Bessel function of the first kind to accurately capture the spatial spillover effect of channel correlation under pre-

diction miss scenarios. Leveraging this analytical framework, we formulate a non-convex optimization problem that captures the dynamic tradeoffs among motion timeliness, pilot overhead, and spatial diversity gain under stringent task deadline constraints.

3) To meet the stringent real-time communication requirements of mobile agents, we design an efficient computational framework combining offline spatial expectation decoupling with online discrete zeroth-order (ZO) gradient tracking. The offline stage decouples the computationally intensive multi-path channel statistical integration from real-time motion states, translating core expectation parameters into constant-time algebraic lookups. The online stage utilizes a discrete ZO gradient ascent engine equipped with heuristic initialization and early stopping criteria. This two-stage design enables microsecond-level adaptive tracking of the optimal window size, with an online time complexity entirely decoupled from the massive number of antenna ports.

2 System Model

2.1 Physical Scenario and Mobility Modeling

Consider a single-antenna base station serving an embodied intelligent agent performing environmental inspection tasks, such as an unmanned surface vehicle (USV). The agent is required to travel from the starting point to the destination within a specified task deadline T_{deadline} , and the total straight-line physical travel distance is denoted as D . During task execution, the instantaneous moving speed of the agent is defined as a continuous variable v .

As the agent travels through the environment, its relative motion and variations of surrounding scatterers induce Doppler shift, resulting in rapid fading of the wireless channel. Let f_c denote the carrier frequency of the system, c denote the speed of light, and f_{env} denote the inherent baseline frequency shift of the surrounding environment. The term $f_d = f_c v/c$ represents the maximum geometric Doppler shift caused by the high-speed mechanical movement of the embodied intelligent agent. The channel coherence time T_c of the system is modeled as:

$$T_c = \frac{1}{f_{\text{env}} + \frac{f_c v}{c}} \quad (1).$$

This model indicates that the faster the moving speed v of the agent, the stronger the Doppler effect, and the shorter the channel coherence time T_c ^[20]. To quantitatively analyze the boundary characteristics of this mobility model, this paper discusses two extreme asymptotic scenarios as follows.

- **Quasi-static boundary:** When the moving speed of the agent node approaches zero ($v \rightarrow 0$) or the agent travels at an extremely low speed, the term $f_d \rightarrow 0$, and the channel coherence time degenerates to $T_c \approx 1/f_{\text{env}}$. At this point, the wireless channel is mainly dominated by macroscopic environmental scatter-

ing, exhibiting the traditional block fading characteristic with relatively sufficient temporal coherence.

- **High-Doppler regime:** When the agent has to travel at full speed ($v \rightarrow v_{\text{max}}$) to meet the tight task deadline, i.e., extremely small T_{deadline} , the dominant term in the denominator evolves into the self-induced Doppler shift f_d . In this case, the Doppler spread expands dramatically, causing the channel coherence time T_c to shrink sharply and non-linearly.

To ensure coherent transmission in highly dynamic environments, the system specifies that the length of each data transmission cycle must be equal to the current channel coherence time $T_c(v)$.

2.2 Multipath Channel and Physical Gain Modeling for FAS

As shown in Fig. 1, the agent node is equipped with a one-dimensional linear FAS. To accurately capture the channel correlation of FAS within an extremely small space, this paper adopts the Sum of Sinusoids (SOS) method to physically model the spatial channel^[21–22].

Assume that radio waves, after being scattered by the surrounding environment, form M independent multipath rays arriving at the agent node. Let the m -th ray, $m \in \{1, 2, \dots, M\}$, have a random angle of arrival $\theta_m \in [0, 2\pi]$ following a uniform distribution, and a random initial phase $\phi_m \in [0, 2\pi]$ following a uniform distribution. Let the total physical length of the fluid antenna be W times the signal wavelength λ . Within this fixed physical length, N virtual communication ports for antenna sliding are uniformly distributed. Taking one end of the antenna as the origin of the physical coordinate system, the precise physical spatial coordinate x_n of the n -th virtual port, $n \in \{1, 2, \dots, N\}$, is expressed as:

$$x_n = \frac{n-1}{N-1} W\lambda \quad (2).$$

Due to the tiny differences in spatial coordinates of each port,

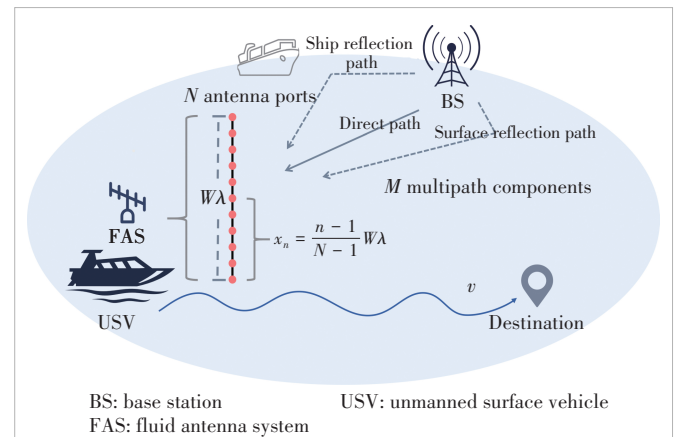


Figure 1. System Diagram of an FAS-enabled agent served by a base station in a multipath environment

continuous phase shifts occur when multipath rays arrive at different ports. The complex channel state h_n received at the n -th virtual port is the superposition of M rays:

$$h_n = \frac{1}{\sqrt{M}} \sum_{m=1}^M \exp \left(j \left(2\pi \frac{x_n}{\lambda} \cos(\theta_m) + \phi_m \right) \right) \quad (3)$$

where $1/\sqrt{M}$ is the normalization coefficient to ensure the conservation of average channel power. The physical channel power gain of the n -th virtual port is defined as:

$$g_n = |h_n|^2 \quad (4)$$

As x_n varies continuously, the spatial channel formed by the gain set $\{g_1, g_2, \dots, g_N\}$ exhibits a continuously fluctuating correlated fading characteristic. There exists a global optimal channel gain over the entire antenna aperture, denoted as:

$$g_{\max} = \max_{n \in \{1, \dots, N\}} g_n \quad (5)$$

To quantitatively characterize the inherent non-independence of FAS at the spatial hardware level, the spatial correlation coefficient between any two virtual ports n and n' is defined as $\rho_{n,n'}$. In the classic rich scattering environment with isotropic scattering, according to Clarke's channel model, the spatial correlation between any two points can be obtained from the zeroth-order Bessel function of the first kind^[4]:

$$\rho_{n,n'} = J_0 \left(\frac{2\pi |x_n - x_{n'}|}{\lambda} \right) \quad (6)$$

Substituting the port physical coordinate formula into the above equation, the spatial correlation between ports can be expressed as a function of the discrete index difference:

$$\rho_{n,n'} = J_0 \left(2\pi W \frac{|n - n'|}{N - 1} \right) \quad (7)$$

The above spatial correlation model demonstrates that the spatial channel formed by the gain set $\{g_1, g_2, \dots, g_N\}$ is not spatially independent across all positions, but exhibits a continuous undulating correlated fading characteristic in space. When the index difference $|n - n'|$ is small, ports maintain a strong positive spatial correlation. This physical property serves as the theoretical cornerstone of the mechanism proposed in this paper: If there exists a minor spatial deviation between the optimal port predicted by the high-level embodied intelligence algorithm and the true global optimal port $g_{\max}$, constrained by the envelope of the correlation coefficient $\rho_{n,n'}$,

opening a local window with K consecutive ports around the predicted port for adaptive sliding sounding can still achieve channel gain close to the global optimum with a very high probability. On the contrary, if ports are mutually independent, any small deviation of the high-level prediction would result in complete detection failure.

2.3 Embodied Artificial Intelligence (EAI) Prediction and Channel Sounding Overhead Model

Acquiring channel state information (CSI) consumes time-domain resources. Let the pilot overhead duration for testing a single FAS port be a constant τ_0 . To obtain the global optimal gain $g_{\max}$, the traditional full scanning mechanism traverses all N ports across the entire aperture, resulting in a delay overhead of $N\tau_0$.

In the proposed mechanism, the embodied intelligent agent leverages the perception model to predict the local interval where the optimal channel is most likely to exist, and instructs the FAS to perform sounding only within a local search window containing K consecutive virtual ports, where the discrete control variable K satisfies $1 \leq K \leq N$. Under this mechanism, the pilot sounding duration per cycle is reduced to $K\tau_0$. Within one coherence time T_c , after deducting the pilot overhead, the time ratio factor $\eta(v, K)$ available for effective service data transmission is modeled as:

$$\eta(v, K) = \max \left[0, 1 - \frac{K\tau_0}{T_c} \right] = \max \left[0, 1 - K\tau_0 \left(f_{\text{env}} + \frac{f_c v}{c} \right) \right] \quad (8)$$

As shown in Fig. 2, this formula reveals the core tradeoff in highly dynamic scenarios that either an excessively large window K or an excessively high speed v will cause the effective transmission time to approach zero. Among them, the perceptual computation phase with negligible time overhead consists of two sub-processes: The EAI perception model outputs the predicted optimal port, and the online ZO optimization algorithm determines the optimal search window size.

Due to the uncertainty in the embodied intelligence's prediction of the environment, this paper defines the probability that

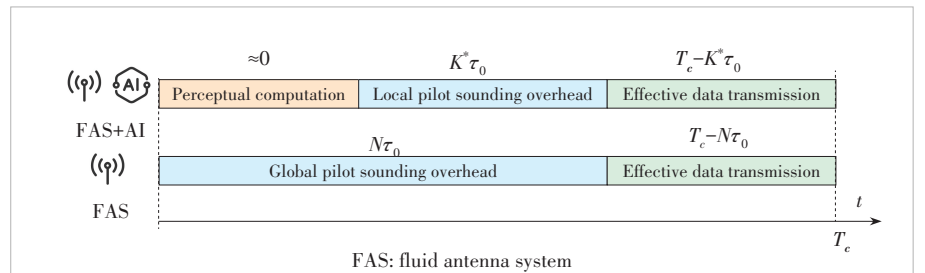


Figure 2. Timing diagram of pilot overhead and effective data transmission within one channel coherence time T_c , comparing the proposed prediction-aided local sounding (FAS+AI) with conventional full-port scanning (FAS)

the embodied intelligence accurately includes the global optimal port within a window of size K as the hit probability $p(K)$. A nonlinear bounded growth function is introduced to characterize this behavior:

$$p(K) = 1 - (1 - p_0) \left(1 - \frac{K-1}{N-1}\right)^\alpha \quad (9).$$

In the formula, the constant $p_0 \in (0,1)$ denotes the baseline perception hit rate under a minimal window ($K=1$); the constant $\alpha \geq 1$ is the confidence index that reflects the performance decay rate of the embodied perception model. This hit probability model satisfies the following boundary characteristics of fluid antennas:

- Single-point baseline: When the system relies entirely on the EAI's single-point prediction ($K=1$), the function rigorously degrades to $p(1) = p_0$, representing the inherent baseline perception confidence of the prediction model.
- Full-aperture scan: When the search window expands to traverse the entire physical antenna aperture ($K=N$), the system invariably captures the global optimum, correctly yielding $p(N) = 1$.

The exponent $\alpha > 1$ serves as a crucial non-linear shape parameter. It mathematically maps to the precision metric (e.g., the scale parameter λ of the Laplace distribution) of the underlying error distribution, characterizing the diminishing marginal probability gain as the search window expands toward the aperture boundaries.

2.4 Hit Probability and Expected Transmission Capacity Model

In summary, when the FAS performs sliding sounding within a window of size K , the system communication presents two mutually exclusive statistical states as follows.

- Prediction hit scenario ($p(K)$): The FAS performs sliding sounding within window K and successfully captures the global optimal gain; that is, the effective gain satisfies $g_{\text{hit}} = g_{\text{max}}$.
- Prediction miss scenario ($1 - p(K)$): The global optimal port lies outside the search window, and the FAS can only obtain a suboptimal instantaneous maximum gain in the erroneous local region, denoted as $\tilde{g}_{\text{local}}(K) = \max_{n \in K_{\text{miss}}} g_n$, where K_{miss} represents the set of consecutive virtual port indices that exclude the true global optimal port.

Based on the microscopic spatial correlation theory derived above, the local extremum $\tilde{g}_{\text{local}}(K)$ in the miss scenario is not spatially independent of the global optimal point g_{max} ; instead, it is constrained by the physical envelope of the spatial correlation coefficient. Furthermore, the parametric multipath model for fluid antennas proposed by Wong et al.^[3] demonstrates that the correlated channel states across the entire FAS aperture can be decoupled into a linear superposition of a global baseline state and local independent fluctuations in the complex voltage do-

main. To overcome the computational bottleneck in calculating the joint expectation of non-independent multivariate Rayleigh distributions and to quantitatively characterize this spatial spillover effect, this paper constructs an asymptotic bridging function. It treats the uncaptured gain $\tilde{g}_{\text{local}}(K)$ within the sounding window as an independent fluctuation, combining it with the global state weighted by the squared Bessel factor $\rho_{\text{eff}}(K)^2$, which maps the expected minimum physical distance into the power domain. Consequently, the conditional effective instantaneous gain in the miss scenario is modeled as a continuous mixed variable that evolves with the window size K :

$$\tilde{g}_{\text{miss}}(K) = \left| \rho_{\text{eff}}(K) \right|^2 g_{\text{max}} + \left(1 - \left| \rho_{\text{eff}}(K) \right|^2\right) \tilde{g}_{\text{local}}(K) \quad (10),$$

where $\rho_{\text{eff}}(K)$ denotes the equivalent spatial correlation factor between the window edge and the true optimal port in the miss scenario. Under the miss event, the true optimal port is assumed to follow a uniform distribution among the remaining $N-K$ ports outside the window. The expected index distance from the optimal port to the nearest boundary of the sounding window can be quantitatively expressed as $\mathbb{E}[d_{\text{min}}] \approx \frac{N-K}{4(N-1)} W\lambda(1 - p_0)$. Accordingly, the equivalent spatial correlation factor $\rho_{\text{eff}}(K)$ can be directly derived from the zeroth-order Bessel function of the first kind:

$$\rho_{\text{eff}}(K) = \max \left(0, J_0 \left(2\pi W \frac{N-K}{4(N-1)} (1 - p_0) \right) \right) \quad (11).$$

Some recent cutting-edge studies have derived closed-form approximations for the average spatial correlation across the entire antenna aperture to break through the integral calculation bottleneck of multipoint joint distributions^[19]. However, such macroscopic averaging algorithms are mainly applied to evaluate the global ergodic capacity of FAS. In the embodied intelligence-assisted architecture proposed in this paper, the underlying FAS performs local sliding sounding based on the dynamic window K . Directly adopting the global static average correlation parameters would eliminate the microscopic distance evolution characteristics between the local window edge and the optimal point. Therefore, to accurately capture the high-resolution spatial dynamics under the local search mechanism, the proposed bridging model does not adopt macroscopic simplification algorithms, but retains the precise physical mapping of the Bessel function.

This model conforms to the underlying physical limits of FAS: When the window is extremely small, i.e., $K \rightarrow 1$, a prediction miss will most likely leave the erroneous single-point sounding far from the optimal solution, and the correlation factor will drop to a low level, resulting in severe attenuation of $\tilde{g}_{\text{miss}}(K)$. As the window gradually expands, i.e., $K \rightarrow N$, even if the high-level algorithm still judges it as a "miss", the physi-

cal edge of the large window statistically approaches the true optimal port continuously, i.e., $\mathbb{E}[d_{\min}] \rightarrow 0$, and the spatial correlation recovers rapidly, $\rho_{\text{eff}} \rightarrow 1$, making $\tilde{g}_{\text{miss}}(K)$ approach g_{max} .

When calculating the system transmission capacity, the dual physical fading from macroscopic distance and microscopic multipath propagation needs to be considered. Since the macroscopic physical displacement of the embodied intelligent agent is negligible relative to the distance to the base station within the extremely short cycle (millisecond level) of local sounding and transmission, this paper assumes that the large-scale path loss dominated by absolute distance remains quasi-static within each local cycle, and models it as a constant local baseline received Signal-to-Noise Ratio (SNR), denoted as γ .

On top of this baseline, the small-scale instantaneous physical gain captured by the fluid antenna sliding within the tiny physical aperture, i.e., g_{max} or $g_{\text{miss}}(K)$, is the real cause of dramatic fluctuations in communication quality. Therefore, the full instantaneous received SNR of the system can be decoupled into the product of the baseline constant and the instantaneous gain. Combined with Shannon's theorem, given the system channel bandwidth B and baseline SNR γ , under a given strategy (v, K) , the instantaneous effective transmission capacity $C_{\text{instant}}(v, K)$ under a single random multipath realization is expressed as:

$$C_{\text{instant}}(v, K) = \eta(v, K) \cdot B \cdot [p(K) \log_2(1 + \gamma g_{\text{max}}) + (1 - p(K)) \log_2(1 + \gamma \tilde{g}_{\text{miss}}(K))] \quad (12).$$

The long-term average expected effective throughput $R(v, K)$ within one communication cycle is modeled as:

$$R(v, K) = \eta(v, K) \cdot B \cdot \mathbb{E}_h \left[p(K) \log_2(1 + \gamma g_{\text{max}}) + (1 - p(K)) \log_2(1 + \gamma \tilde{g}_{\text{miss}}(K)) \right] \quad (13),$$

$$R(v, K) = \eta(v, K) \cdot B \cdot [p(K) \bar{C}_{\text{hit}} + (1 - p(K)) \bar{C}_{\text{miss}}(K)] \quad (14),$$

where $\mathbb{E}_h[\cdot]$ denotes the mathematical expectation over all possible realizations of the continuous spatial multipath channel.

2.5 System Extension for 2D FAS

The above work constructs a spatial bridging model for FAS based on one-dimensional linear topology. As the demand for spatial degrees of freedom within a physical aperture rises, two-dimensional planar fluid antennas (Planar FAS) have emerged as a key evolutionary direction for terminal antenna design. Recent studies demonstrate that two-dimensional fluid antennas deliver superior geometric packing performance and spatial diversity gain under a constrained aperture^[23], and have been deployed in sophisticated network architectures such as simultaneous wireless information and power transfer (SWIPT) and feder-

ated learning^[24]. To verify the scalability of the theoretical framework proposed in this paper, this section extends the one-dimensional baseline model to a two-dimensional planar space.

In the two-dimensional extended model (Fig. 3), the physical panel of the fluid antenna is defined as a rectangular region with dimensions $W_y \lambda \times W_z \lambda$, on which $N_y \times N_z = N_{\text{total}}$ discrete virtual ports are uniformly arranged. Accordingly, the single spatial index is extended to a two-dimensional coordinate pair (m, n) . Under the assumption of isotropic scattering, the microscopic spatial correlation coefficient $\rho_{(m, n), (m', n')}$ between any two ports still follows the zeroth-order Bessel function of the first kind, but its argument is replaced with the two-dimensional Euclidean distance:

$$d_{(m, n), (m', n')} = \lambda \sqrt{\left(\frac{m - m'}{N_y - 1} W_y\right)^2 + \left(\frac{n - n'}{N_z - 1} W_z\right)^2} \quad (15).$$

When the underlying hardware is extended to a two-dimensional panel, the local detection window output by the upper-layer perception model changes from a one-dimensional line segment to a two-dimensional rectangular search box containing K ports. For the prediction miss state, solving the expected minimum physical distance $\mathbb{E}[d_{\min}]$ from the unknown global optimal port to the edge of the search box involves complex two-dimensional boundary integration. Existing literature usually adopts multivariate distribution tools such as Copula theory for accurate planar correlation modeling^[25], but the high analytical complexity of Copula-based methods is not applicable to the cross-layer joint optimization problem in this paper. Therefore, the equivalent area method is adopted for dimensionality reduction and approximate analysis. The total area of the two-dimensional panel is $A_{\text{total}} = W_y W_z \lambda^2$, and the equivalent to-

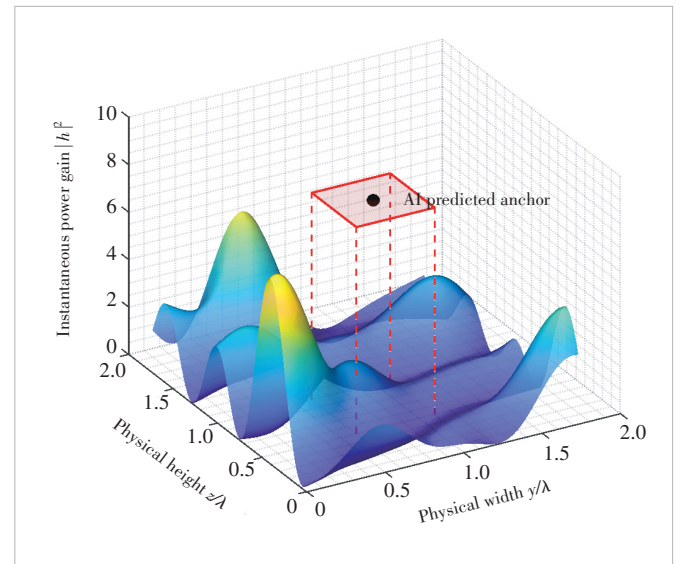


Figure 3. Schematic diagram of spatial relationship for 2D FAS

tal side length is $D_{\text{total}} = \sqrt{W_y W_z} \lambda$. The area of the search box with K ports is $A_K = \frac{K}{N_{\text{total}}} W_y W_z \lambda^2$, and the equivalent inner side length is $D_K = \sqrt{A_K} = \sqrt{\frac{K}{N_{\text{total}}}} \sqrt{W_y W_z} \lambda$. When the remaining length is $D_{\text{total}} - D_K$, the expected distance from a uniformly distributed random point to the nearest edge is $\frac{1}{4} (D_{\text{total}} - D_K)$. The derived two-dimensional expected minimum physical distance $\mathbb{E}[d_{\text{min}}^{2D}]$ is approximated as:

$$\mathbb{E}[d_{\text{min}}^{2D}] \approx \frac{\sqrt{W_y W_z}}{4} \left(1 - \sqrt{\frac{K}{N_{\text{total}}}} \right) \lambda \quad (16).$$

Substituting this expected distance into the correlation envelope function yields the equivalent spatial correlation factor in the two-dimensional space:

$$\rho_{\text{eff}}^{2D}(K) = \max \left(0, J_0 \left(\frac{2\pi \mathbb{E}[d_{\text{min}}^{2D}]}{\lambda} \right) \right) \quad (17).$$

Furthermore, the equivalent suboptimal gain under two-dimensional conditions can also be expressed in a linear bridging form:

$$\tilde{g}_{\text{miss}}(K) = |\rho_{\text{eff}}^{2D}(K)|^2 g_{\text{max}} + \left(1 - |\rho_{\text{eff}}^{2D}(K)|^2 \right) \tilde{g}_{\text{local}}(K) \quad (18).$$

Via mapping based on the equivalent area method, the physical dimensions of the antenna are mathematically unified and abstracted into a single discrete resource variable K , which denotes the total number of ports within the search window. Accordingly, the expected throughput objective function $R(v, K)$ formulated earlier and the subsequent two-stage offline-online algorithm are formally compatible with one-dimensional and two-dimensional FAS via this approximation.

It should be noted that the above treatment is a dimension-reduction approximation. It neglects subtle physical characteristics, including the azimuth offset of the two-dimensional rectangular window, abrupt corner gain variations, and differences in the planar distribution of multipath arrival angles, leading to a precision gap compared with the rigorous model for two-dimensional planar FAS. Full-dimensional precise modeling for two-dimensional topologies, optimization of the aspect ratio and shape of rectangular windows, and dedicated joint scheduling algorithms tailored to two-dimensional spaces can be systematically investigated as independent research directions in future work.

3 Joint Optimization Problem Formulation

To maximize the communication efficiency of embodied intelligent agents during the execution of highly dynamic tasks, this

paper focuses on optimizing the underlying fluid antenna adaptive sensing mechanism. In practical operational scenarios for embodied intelligence, the agent's instantaneous moving speed v is strictly governed by high-level mechanical navigation and motion planning. Meanwhile, the prediction accuracy p_0 of the environmental perception model fluctuates in real time due to physical conditions such as local illumination and occlusions.

Therefore, from the perspective of the physical-layer communication link, the variable pair (v, p_0) is not a decision-control variable that can be actively tuned by the communication system; instead, it represents instantaneous state variables determined by the external environment and task requirements. Under such rapidly time-varying states, the core decision-making authority of the communication physical layer lies in flexibly adjusting the local search window size K of the FAS. Given the system channel bandwidth B and local average received SNR γ , the long-term expected effective throughput over a single communication cycle under the current instantaneous state (v, p_0) is defined as the objective function. The joint optimization problem P1 is formulated as follows:

$$\begin{aligned} \text{P1: } \max_K \quad & R(K|v, p_0) = \eta(v, K) \cdot B \cdot \\ & [p(K)\bar{C}_{\text{hit}} + (1 - p(K))\bar{C}_{\text{miss}}(K)] \\ \text{s.t.} \quad & \text{C1: } K\tau_0 < T_c(v) \\ & \text{C2: } K \in \{1, 2, \dots, N\} \end{aligned} \quad (19).$$

3.1 Offline Decoupling of Spatial Channel Expectation

The objective function consists of two components. The constant $\bar{C}_{\text{hit}}$ denotes the expected channel capacity when the globally optimal antenna port is successfully captured within the search window. The variable $\bar{C}_{\text{miss}}(K)$ denotes the conditional expected capacity under prediction miss, which is bounded by the Bessel-based spatial correlation envelope and increases with K as the window edge approaches the true optimal port. The time-ratio factor $\eta(v, K) = 1 - \tau_p K / T_c(v)$ characterizes the fraction of the coherence time remaining for data transmission after pilot sounding.

Constraint C1 is the timing constraint, which guarantees that the pilot sounding overhead is strictly less than the current channel coherence time. Constraint C2 restricts the search window to a positive integer that does not exceed the total number of physical antenna ports N .

3.2 Coupling Structure of the Objective Function

Problem P1 is a one-dimensional discrete nonlinear programming problem subject to hard timing constraints. Although the decision variable is a single scalar K , the objective function exhibits a coupled structure in which three physical mechanisms evolve with K in opposing directions:

- Time-ratio factor $\eta(v, K)$: monotonically decreasing with K , since window expansion increases pilot time consumption and

reduces the time available for effective data transmission.

- Sensing hit probability $p(K)$: monotonically increasing with K , since a larger window statistically raises the probability of covering the true optimal port.
- Miss-state expected capacity $\bar{C}_{\text{miss}}(K)$: monotonically increasing with K , since window expansion shortens the physical distance between the search boundary and the true optimal port, thereby elevating the equivalent Bessel spatial correlation factor and the local suboptimal gain.

The product of the decreasing time-ratio factor and the increasing spatial gain term creates a competing structure that generally favors an intermediate window size. However, because the spatial gain term further involves the nonlinear bounded hit probability function $p(K)$, the Bessel-based correlation factor in $\bar{C}_{\text{miss}}(K)$, and Monte-Carlo-smoothed statistical expectations, its precise shape depends on system parameters and may exhibit local fluctuations. A rigorous analytical proof of global unimodality or quasi-concavity is therefore not available. Numerical experiments across the parameter range considered in this study consistently indicate a predominantly unimodal trend (i.e., the throughput first increases and then decreases with K), which provides a heuristic basis for the gradient-based search algorithm developed in Section 4. The algorithm further mitigates the risk of local irregularities through heuristic anchor initialization and early termination.

Since the underlying integral operator of the objective function has no explicit closed-form analytical solution, and the environmental state (v, σ) switches frequently in real time, directly evaluating all candidate windows online induces severe latency. This motivates the low-complexity algorithm design in the following section.

4 Algorithm Design

To address the non-convex nature of the optimization problem and the stringent constraint of ultra-short coherence time for embodied intelligent agents, this paper proposes a two-stage algorithm framework termed “offline spatial expectation decoupling and online discrete zeroth-order tracking”. This framework offloads computationally intensive electromagnetic statistical integration to the offline phase, while adopting a discrete ZO gradient algorithm to accomplish adaptive optimization with minimal computational overhead during the online phase.

4.1 Offline Decoupling of Spatial Channel Expectation

Based on the ergodicity assumption of wireless channels, the spatial fading probability density function of multipath channels follows a time-invariant distribution within a broad class of physical scenarios (e.g., rich-scattering environments). Accordingly, the intricate statistical integration operations can be fully decoupled from real-time motion states.

Offline precomputation of the algorithm is implemented on a ground server. Given the system physical parameters and reference SNR γ , we generate N_{sim} sample realizations of Clarke’s

isotropic scattering multipath channel, then compute the expected hit upper bound $\bar{C}_{\text{hit}}$ corresponding to the global extreme point of the system. Next, we traverse all candidate search windows $K \in \{1, \dots, N\}$, and compute the array of bridged expected capacities under miss conditions $\bar{C}_{\text{miss}}(K)$ with the equivalent Bessel spatial correlation factor incorporated. To eliminate spurious numerical peaks (ripples) induced by finite Monte Carlo sampling, a moving average filter is applied to the one-dimensional array $\bar{C}_{\text{miss}}(K)$ to restore smoothness.

Upon completion of offline computation, the system only needs to transmit one scalar value and a one-dimensional array of length N to the on-board memory of the agent. In this way, the integral operator $\mathbb{E}_h[\cdot]$ embedded in the objective function is decoupled and converted into an algebraic memory lookup with $\mathcal{O}(1)$ computational complexity.

4.2 Online Algorithm Based on Discrete ZO Gradient

During task execution by the agent, high-frequency fluctuations of the instantaneous state (v, p_0) cause the unimodal optimal point K^* of the objective function to drift continuously. This algorithm adopts the discrete ZO gradient ascent method to realize microsecond-scale adaptive tracking without exhaustive global search and state-space quantization errors inherent to conventional large lookup tables.

Prior anchor initialization: A heuristic starting point is constructed based on perception confidence. When p_0 is large, the optimal solution tends to correspond to a small search window. The initial point is defined as:

$$K_{\text{curr}} = \max\left(1, \lfloor N \cdot (1 - p_0) \cdot \beta \rfloor\right) \quad (20),$$

where β denotes a tuning constant.

ZO two-sided difference estimation: At the current iterate K_{curr} , single-step perturbations are applied to its left and right discrete neighbors. The lightweight objective function is evaluated to obtain $R(K_{\text{curr}} - 1)$, $R(K_{\text{curr}})$, and $R(K_{\text{curr}} + 1)$. The discrete gradient is approximated via central difference:

$$\hat{\nabla}_K R = \frac{R(K_{\text{curr}} + 1) - R(K_{\text{curr}} - 1)}{2} \quad (21).$$

Adaptive step sizing and early termination: The estimated gradient is normalized adaptively in magnitude to guide the search window toward the extremum with dynamic step sizes. Benefiting from the unimodal property of the objective function, the optimization process is considered to have reached the global maximum once $R(K_{\text{curr}}) \geq R(K_{\text{curr}} - 1)$ and $R(K_{\text{curr}}) \geq R(K_{\text{curr}} + 1)$. A hard stopping criterion is then activated to output the optimal K^* .

The execution logic of the online phase is summarized in Algorithm 1.

Algorithm 1. Online Adaptive Tracking via Discrete ZO Gradient Ascent

Input: Instantaneous state vector (v, p_0) , pre-computed offline constant $\bar{C}_{\text{hit}}$, pre-computed offline array $\bar{C}_{\text{miss}}[1 \cdots N]$, learning rate μ , tuning constant β .

```

1: Calculate coherence time bound:  $T_c = 1/(f_{\text{env}} + f_c v/c)$ 
2: Calculate maximum permissible window constraint:
    $K_{\text{max}} = \lfloor T_c / \tau_0 \rfloor$ 
3: if  $K_{\text{max}} < 1$  then
4:   return Exception (Velocity exceeds physical tracking limit)
5: end if
6: Initialize heuristic anchor:
    $K_{\text{curr}} = \max\left(1, \min\left(K_{\text{max}}, \lfloor N \cdot (1 - p_0) \cdot \beta \rfloor\right)\right)$ 
7: for iter = 1 to MaxIter do
8:   Define discrete neighborhood:
      $K_{\text{left}} = \max(1, K_{\text{curr}} - 1)$ ,  $K_{\text{right}} = \min(K_{\text{max}}, K_{\text{curr}} + 1)$ 
9:   Evaluate  $R(K_{\text{left}})$ ,  $R(K_{\text{curr}})$ , and  $R(K_{\text{right}})$  using
     decoupled algebraic components
10:  if  $R(K_{\text{curr}}) \geq R(K_{\text{left}})$  and  $R(K_{\text{curr}}) \geq R(K_{\text{right}})$  then
11:     $K^* = K_{\text{curr}}$ 
12:    break
13:  end if
14:  Estimate discrete central difference gradient:
15:    if  $K_{\text{curr}} == 1$  then  $\hat{\nabla}_K R = R(K_{\text{right}}) - R(K_{\text{curr}})$ .
16:    else if  $K_{\text{curr}} == K_{\text{max}}$  then  $\hat{\nabla}_K R = R(K_{\text{curr}}) - R(K_{\text{left}})$ 
17:    else  $\hat{\nabla}_K R = (R(K_{\text{right}}) - R(K_{\text{left}}))/2$ 
18:  end if
19:  Compute adaptive step size: step = round( $\mu \cdot \hat{\nabla}_K R_{\text{norm}}$ )
20:  Update current window:
      $K_{\text{curr}} = \max\left(1, \min\left(K_{\text{max}}, K_{\text{curr}} + \text{step}\right)\right)$ 
21: end for
Output: Optimal search window size  $K^*$ 

```

4.3 Time Complexity and Real-Time Performance Analysis

To verify the feasibility of the proposed framework under the hard real-time communication constraints of embodied intelligence, this section conducts an asymptotic complexity analysis from both temporal and spatial dimensions.

1) Spatial complexity analysis

In terms of storage overhead, the system decouples complex physical integration from macroscopic environmental states, eliminating the need for the agent's on-board terminal to store massive multi-dimensional state grids required by the conventional lookup table (LUT) method. The online optimization only permanently retains one constant of expected upper bound $\bar{C}_{\text{hit}}$ and a one-dimensional expected array $\bar{C}_{\text{miss}}[K]$ of length N in memory. Accordingly, the spatial complexity of the system is bounded by $\mathcal{O}(N)$, and its RAM footprint imposes nearly negligible load on on-board tiny NPUs or microcontrollers.

2) Temporal complexity and real-time performance analysis

If the conventional exhaustive global search is adopted, the system must evaluate all N candidate windows within each communication cycle, resulting in an online time complexity linear

in the number of antenna ports, namely $\mathcal{O}(N)$. This approach incurs severe latency when deployed for large-scale fluid antenna systems. By contrast, the discrete ZO tracking framework proposed in this paper achieves a significant reduction in computational overhead:

- Complexity of single function evaluation: Benefiting from offline expectation decoupling, computationally expensive integration and Bessel series calculations are removed from the online objective function. Each single evaluation of $R(K)$ merely involves constant-level algebraic multiplication-addition operations and an $\mathcal{O}(1)$ memory lookup.

- Convergence complexity of optimization iterations: The ZO gradient iteration is hard-constrained within a predefined maximum iteration count I_{max} . Moreover, owing to the inherent unimodal property of the objective function, combined with heuristic anchor initialization and an early stopping mechanism, the actual number of iterations required by the algorithm is generally much smaller than I_{max} .

Hence, the worst-case time complexity of online adaptive optimization is $\mathcal{O}(I_{\text{max}})$. Since I_{max} is a constant bound independent of the antenna scale N , the algorithm achieves complete decoupling from system size in the asymptotic sense. When deployed on mainstream on-board computing hardware, the optimization procedure with $\mathcal{O}(I_{\text{max}})$ complexity runs on a sub-microsecond timescale, which is much shorter than the millisecond-scale coherence time limit $T_c(v)$ induced by the Doppler shift in highly dynamic scenarios. This characteristic fully satisfies the hard real-time communication requirements for high-frequency autonomous operation of embodied intelligent agents.

5 Simulation Results and Analysis

This section validates the performance of the proposed embodied intelligence-aided FAS algorithm. The key parameter settings are specified as follows: carrier frequency $f_c = 2.4$ GHz, total number of antenna ports $N = 100$, antenna length $W = 2$, number of multipath components $M = 20$, pilot time consumption per port τ_0 , and prediction accuracy of the EAI perception model p_0 .

5.1 Analysis of Effective Throughput and Overhead Catastrophe Avoidance

As illustrated in Fig. 4, theoretically, the optimal window size K^* that maximizes the effective throughput R increases as τ_0 , p_0 , or v decreases. The rise in the agent's moving speed v and the growth of single-port pilot overhead τ_0 jointly squeeze the system's available time budget. As shown in the partial enlarged view for $\tau_0 = 0.05$ ms, the throughput R drops to zero at $K = 80$ under $v = 30$ m/s, which corresponds to the overhead catastrophe.

A higher prediction accuracy p_0 eliminates the necessity of selecting K sampling points around the predicted position for pilot transmission to expand fault tolerance. According to the ini-

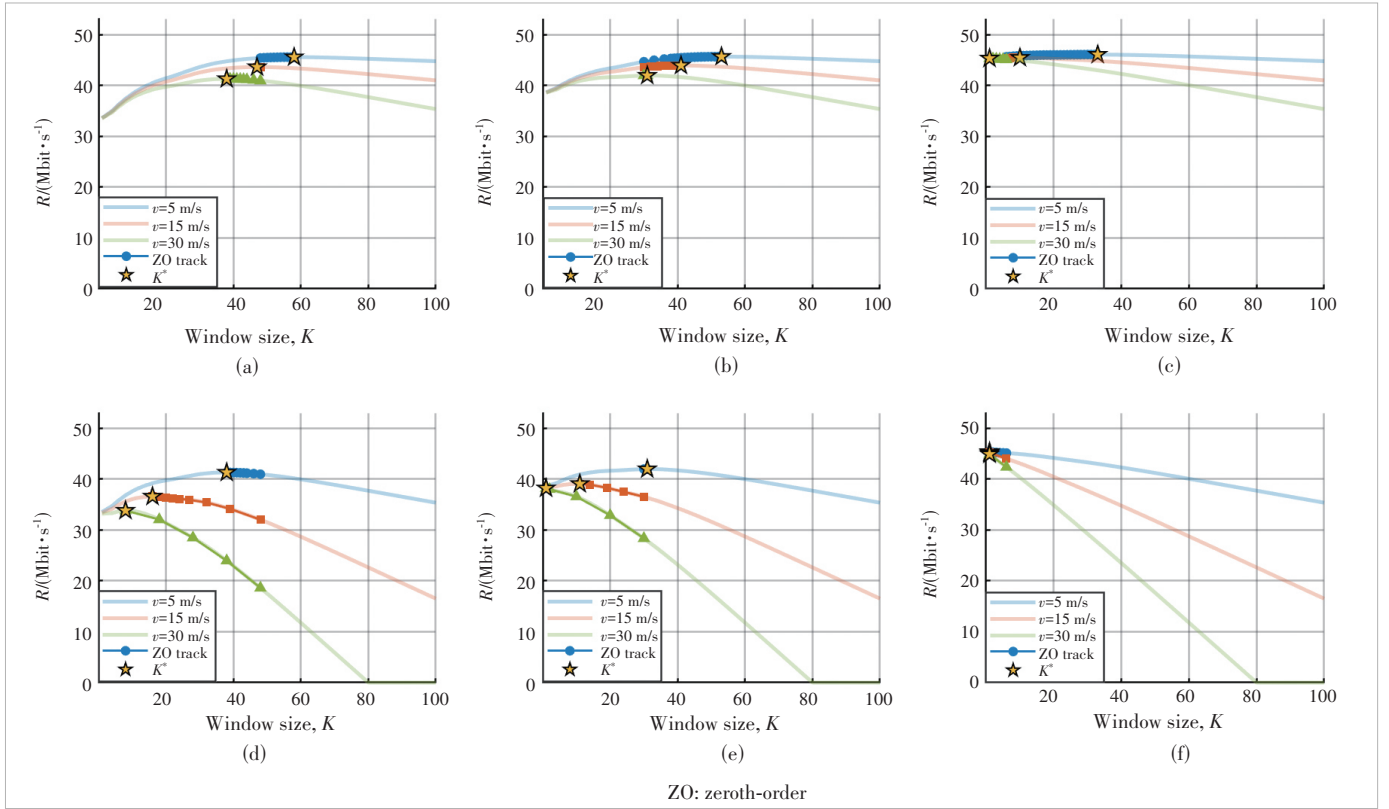


Figure 4. Simulation results of R versus K : (a) $\tau_0 = 0.01$ ms, $p_0 = 0.2$; (b) $\tau_0 = 0.01$ ms, $p_0 = 0.5$; (c) $\tau_0 = 0.01$ ms, $p_0 = 0.9$; (d) $\tau_0 = 0.05$ ms, $p_0 = 0.2$; (e) $\tau_0 = 0.05$ ms, $p_0 = 0.5$; (f) $\tau_0 = 0.05$ ms, $p_0 = 0.9$

tialization formula $K_{\text{curr}} = \max(1, \lfloor N \cdot (1 - p_0) \cdot \beta \rfloor)$ in the proposed algorithm, the initial sensing window size exhibits a negative correlation with the prediction accuracy p_0 of the EAI perception model. Under poor sensing conditions, the algorithm adaptively anchors the starting point to a region with larger K to compensate for sensing degradation by exploiting spatial diversity gains.

5.2 Adaptive Regulation of Optimal Window and Perception Accuracy Gain

Fig. 5 plots the variation of R against v under diverse operating conditions, where the three curves correspond to the proposed algorithm, the fixed-window scheme with $K = 20$, and the exhaustive global search with $K = N$, respectively. Simulation results reveal that in low-mobility scenarios with low pilot overhead, the channel coherence time is sufficiently long. Conventional scanning schemes can attain full spatial diversity gains and deliver decent system performance, exhibiting marginal performance gaps compared with the proposed algorithm. Nevertheless, as v and τ_0 rise, the throughput of traditional approaches drops sharply and eventually falls to zero beyond a certain velocity threshold, namely the overhead catastrophe. In contrast, the algorithm developed in this paper dynamically shrinks the search window K according to the moving velocity. At the cost of a slight loss in maximum spatial diversity gain, it significantly

reduces pilot overhead time, thereby sustaining high and stable throughput even in highly dynamic operating regions.

Fig. 6 reveals the adaptive mechanism of the system. The adaptive window adjustment reflects the tradeoff between spatial domain gain and time domain overhead. When the perception accuracy p_0 is low, the system expands the search window K^* to obtain spatial diversity gain to compensate for sensing degradation. When an increase in moving velocity v leads to temporal compression, the system adaptively shrinks the search window to guarantee valid transmission duration. Under conditions of extremely high moving velocity or extremely high perception accuracy, the system tends to select a smaller optimal window and eventually converges to single-point prediction. This indicates that in high-mobility scenarios, the system must rely on high-precision prediction mechanisms to fundamentally avoid the overhead catastrophe.

5.3 Validation of Spatial Correlation Bridging Model

Fig. 7 verifies the gain boundary of the high-level perception accuracy p_0 of the embodied intelligence model over underlying wireless transmission performance. The sensitivity of throughput to perception accuracy increases significantly with moving velocity. In low-speed scenarios, the dependence of system performance on prediction accuracy is relatively weak. In high-speed scenarios, the channel coherence time is extremely short, and

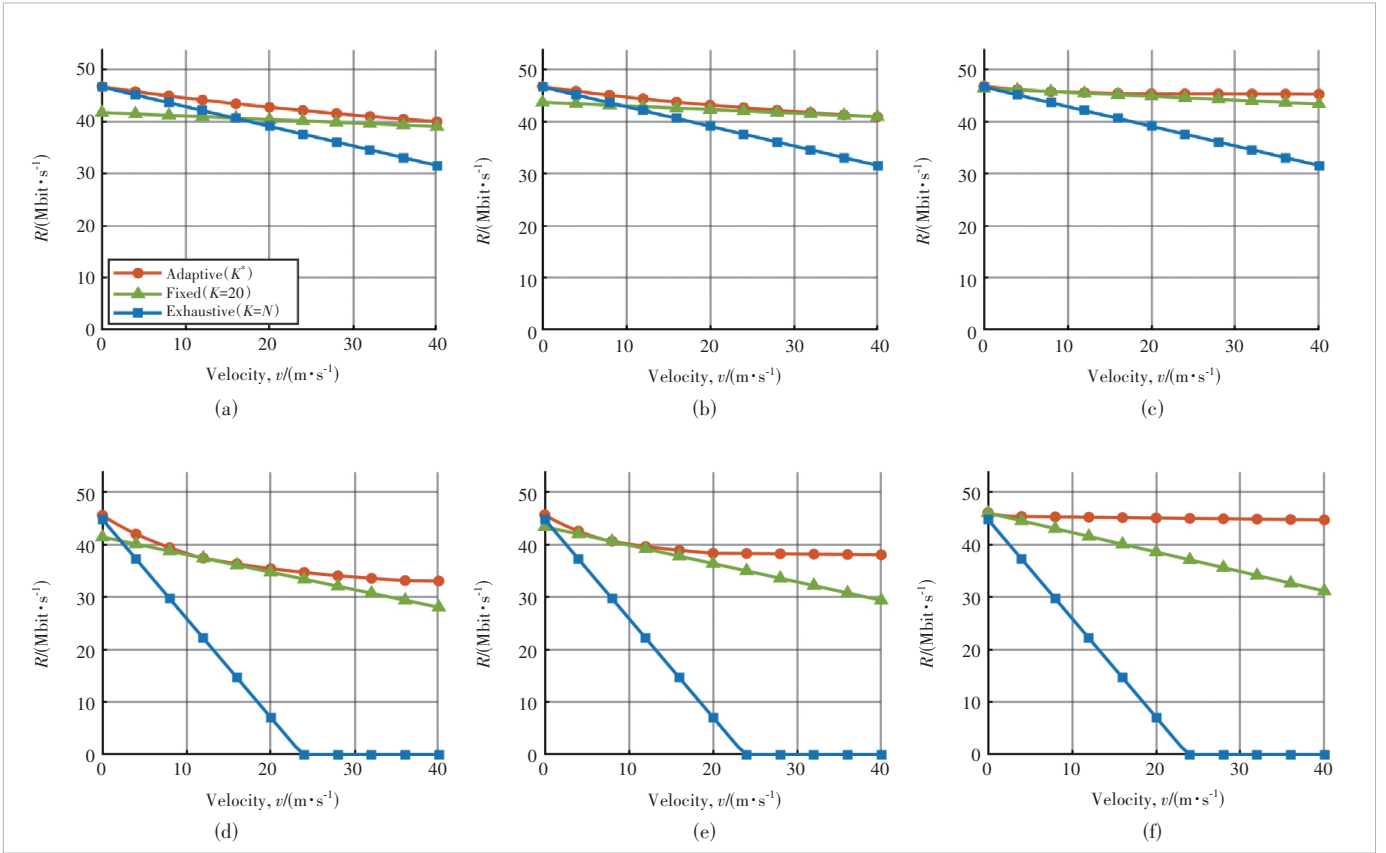


Figure 5. Simulation results of R versus v : (a) $\tau_0 = 0.01$ ms, $p_0 = 0.2$; (b) $\tau_0 = 0.01$ ms, $p_0 = 0.5$; (c) $\tau_0 = 0.01$ ms, $p_0 = 0.9$; (d) $\tau_0 = 0.05$ ms, $p_0 = 0.2$; (e) $\tau_0 = 0.05$ ms, $p_0 = 0.5$; (f) $\tau_0 = 0.05$ ms, $p_0 = 0.9$

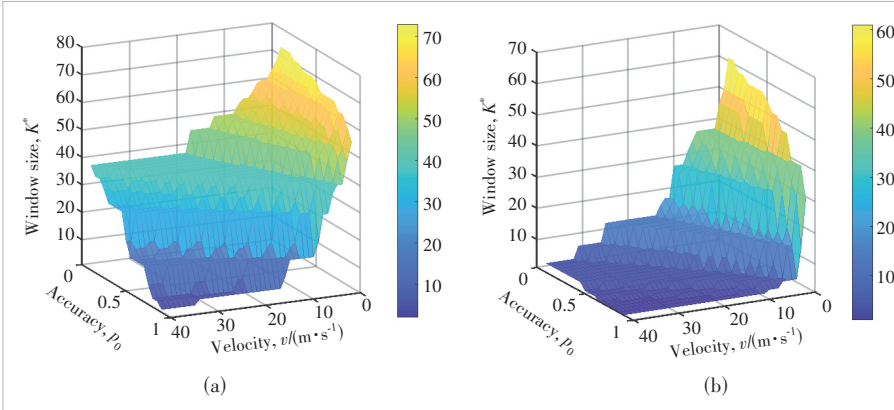


Figure 6. Adaptive adjustment of optimal window size with prediction accuracy and moving velocity: (a) $\tau_0 = 0.01$ ms; (b) $\tau_0 = 0.05$ ms

the system becomes highly sensitive to pilot overhead; improving perception accuracy allows the underlying fluid antenna to complete high-confidence detection within an extremely narrow window, thereby substantially increasing the system throughput. As the perception accuracy continues to improve, the optimal throughputs at different velocities rise steadily and show a convergence trend. This demonstrates that a high-precision embodied perception model can effectively compensate for the physi-

cal transmission deterioration caused by high mobility.

Fig. 8 presents Monte Carlo validation for the theoretical spatial bridging model proposed in Section 2.4. The results reveal that the equivalent suboptimal capacity attainable upon system prediction miss is not a constant. As the local detection window K expands, the physical boundary of the error window gradually converges toward the true optimal port from a macroscopic statistical perspective. Benefiting from the very small port spacing of fluid-state antennas, the underlying Bessel spatial correlation generates a spatial spillover effect, enabling the

conditional miss capacity to converge smoothly and non-linearly to the theoretical upper bound $\bar{C}_{\text{hit}}$. This verifies the physical rigor of the proposed bridging model.

6 Conclusions

To address the overhead catastrophe incurred by full-port scanning of fluid antennas in highly mobile dynamic scenarios,

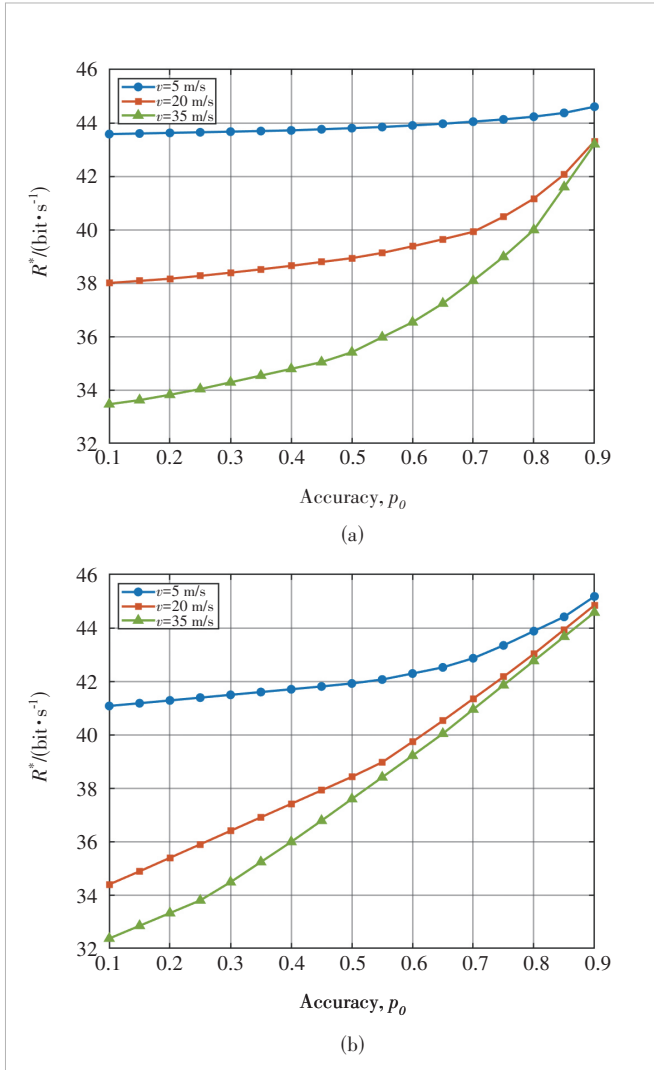


Figure 7. Effective throughput R versus prediction accuracy p_0 under different moving velocities: (a) $\tau_0 = 0.01$ ms; (b) $\tau_0 = 0.05$ ms

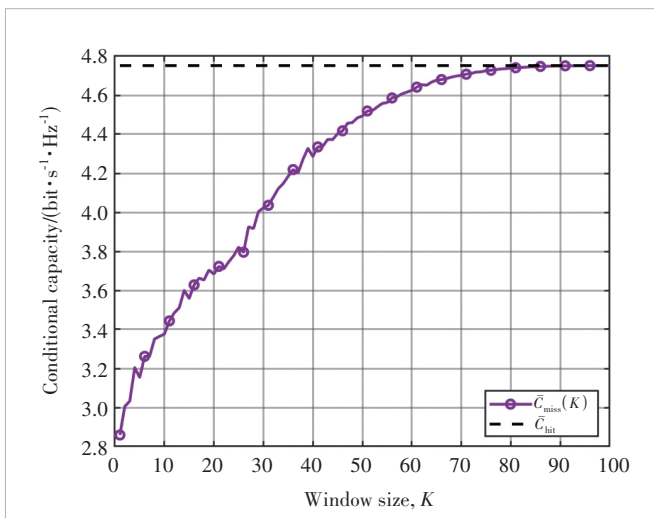


Figure 8. Monte Carlo validation of the spatial bridging model

this paper proposes a predictive error-tolerant window-enhanced adaptive communication mechanism for fluid antennas. Built upon the overhead compression capability of port prediction, the proposed scheme compensates for prediction deviations via a local search window, improves system fault tolerance and channel gain under controllable overhead, and achieves a dynamic tradeoff between overhead and robustness.

This paper establishes a quantitative analysis model for long-term expected throughput. It characterizes the inherent correlation between prediction accuracy and window size via a nonlinear hit probability function, derives a gain bridging model for miss cases based on the zeroth-order Bessel function of the first kind, quantifies the gain overflow effect brought by spatial channel correlation, and formulates a discrete optimization problem subject to timing constraints. The model further reveals the dynamic tradeoffs among moving speed, prediction accuracy, pilot overhead, and spatial diversity gain.

A low-complexity two-stage offline-online solving algorithm is designed in this work. In the offline stage, multipath channel statistical integration is decoupled from real-time motion states, and core expected parameters are precomputed, simplifying online objective function evaluation to constant-time table lookup operations. In the online stage, discrete ZO gradient ascent serves as the core optimization engine. Combined with heuristic initialization and early stopping mechanisms, the algorithm realizes microsecond-scale adaptive tracking of the optimal window without explicit analytical gradients. Its online time complexity is decoupled from the total number of antenna ports, satisfying the hard real-time communication requirements of high-dynamic environments.

Simulation results validate the performance superiority and adaptive mechanism of the proposed scheme: The system dynamically adjusts the window size according to moving speed and prediction accuracy. It fully harvests spatial diversity gains under low-speed conditions and automatically converges to single-point prediction mode at high speeds, fundamentally avoiding overhead catastrophe. Improvements in prediction accuracy can further cut window overhead, and such performance gains are particularly prominent in high-mobility scenarios.

The core model and algorithm developed herein are designed for one-dimensional linear fluid antennas only, while two-dimensional planar scenarios are validated within the framework via dimension-reduction approximation. Future research may extend to precise modeling of two-dimensional fluid antennas, multi-agent joint channel scheduling, and dynamic adaptation of prediction confidence.

References

- [1] Liu H P, Guo D, Cangelosi A. Embodied intelligence: a synergy of morphology, action, perception and learning [J]. *ACM computing surveys*, 2025, 57 (7): 1 – 36. DOI: 10.1145/3717059
- [2] Gao Y, Chen J J, Li D P. Intelligence driven wireless networks in B5G and

- 6G era: a survey [J]. ZTE communications, 2024, 22(3): 99 – 105. DOI: 10.12142/ZTECOM.202403012
- [3] Wong K K, Shojaeifard A, Tong K F, et al. Fluid antenna systems [J]. IEEE transactions on wireless communications, 2021, 20(3): 1950 – 1962. DOI: 10.1109/TWC.2020.3037595
- [4] Wong K K, Tong K F. Fluid antenna multiple access [J]. IEEE transactions on wireless communications, 2022, 21(7): 4801 – 4815. DOI: 10.1109/TWC.2021.3133410
- [5] Wong K K, Shojaeifard A, Tong K F, et al. Performance limits of fluid antenna systems [J]. IEEE communications letters, 2020, 24(11): 2469 – 2472. DOI: 10.1109/LCOMM.2020.3006554
- [6] Wong K K, Chae C B, Tong K F. Compact ultra massive antenna array: a simple open-loop massive connectivity scheme [J]. IEEE transactions on wireless communications, 2024, 23(6): 6279 – 6294. DOI: 10.1109/TWC.2023.3330932
- [7] Xu H, Wong K K, New W K, et al. Capacity maximization for FAS-assisted multiple access channels [J]. IEEE transactions on communications, 2025, 73(7): 4713 – 4731. DOI: 10.1109/TCOMM.2024.3516499
- [8] Zhang Q, Shao M J, Zhang T, et al. An efficient sum-rate maximization algorithm for fluid antenna-assisted ISAC system [J]. IEEE communications letters, 2025, 29(1): 200 – 204. DOI: 10.1109/LCOMM.2024.3510334
- [9] Zhang Z J, Zhu J A, Dai L L, et al. Successive Bayesian reconstructor for channel estimation in fluid antenna systems [J]. IEEE transactions on wireless communications, 2025, 24(3): 1992 – 2006. DOI: 10.1109/TWC.2024.3515135
- [10] Xu B W, Chen Y, Cui Q M, et al. Sparse Bayesian learning-based channel estimation for fluid antenna systems [J]. IEEE wireless communications letters, 2025, 14(2): 325 – 329. DOI: 10.1109/LWC.2024.3500218
- [11] Xiao Z Y, Cao S Q, Zhu L P, et al. Channel estimation for movable antenna communication systems: a framework based on compressed sensing [J]. IEEE transactions on wireless communications, 2024, 23(9): 11814 – 11830. DOI: 10.1109/TWC.2024.3385110
- [12] Tang E Q, Guo W, He H T, et al. Accurate and fast channel estimation for fluid antenna systems with diffusion models [PP/OL]. V1. arXiv (2025-05-08)[2025-08-19]. <https://doi.org/10.48550/arXiv.2505.04930>
- [13] Zhang H B, Wang J L, Wang C, et al. Learning-induced channel extrapolation for fluid antenna systems using asymmetric graph masked autoencoder [J]. IEEE wireless communications letters, 2024, 13(6): 1665 – 1669. DOI: 10.1109/LWC.2024.3386153
- [14] Ji S Y, Psomas C, Thompson J. Correlation-based machine learning techniques for channel estimation with fluid antennas [C]//Proceedings of ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 8891 – 8895. DOI: 10.1109/ICASSP48485.2024.10447499
- [15] Pakravan S, Ahmadzadeh M, Zeng M, et al. Fluid antenna systems under channel uncertainty and hardware impairments: trends, challenges, and future research directions [PP/OL]. V1. arXiv (2026-01-30)[2026-04-01]. <https://doi.org/10.48550/arXiv.2601.22989>
- [16] Deng T S, Gao Y, Zhang T, et al. Integrating large language models into fluid antenna systems: a survey [J]. Sensors, 2025, 25(16): 5177. DOI: 10.3390/s25165177
- [17] Zhang Y L, Yin H F, Li W D, et al. Fluid antenna port prediction based on large language models [C]//Proceedings of 2025 IEEE Global Communications Conference. IEEE, 2025: 4934 – 4939. DOI: 10.1109/GLOBECOM59602.2025.11432485
- [18] Chai Z, Wong K K, Tong K F, et al. Port selection for fluid antenna systems [J]. IEEE communications letters, 2022, 26(5): 1180 – 1184. DOI: 10.1109/LCOMM.2022.3152451
- [19] Wong K K, Tong K F, Chen Y, et al. Closed-form expressions for spatial correlation parameters for performance analysis of fluid antenna systems [J]. Electronics letters, 2022, 58(11): 454 – 457. DOI: 10.1049/ell2.12487
- [20] Shan Y R, Wang F G, Hao Y X, et al. Doppler rate estimation for OTFS via large-scale antenna array [J]. ZTE Communications, 2025, 23(1): 115 – 122. DOI: 10.12142/ZTECOM.202501015
- [21] Liu Zhipeng L K. Dual-polarized 2D beam-scanning antenna based on reconfigurable reflective elements [J]. ZTE communications, 2025, 23(1): 85 – 89. DOI: 10.12142/ZTECOM.202501011
- [22] Lu M Y, Bai L, Han Z R, et al. 6G digital twin enabled channel modeling for Beijing central business district [J]. ZTE Communications, 2025, 23(2): 31 – 45. DOI: 10.12142/ZTECOM.202502005
- [23] Zhang Z T, Xu J Y, Wong K K, et al. Finite-aperture planar fluid antenna array [PP/OL]. V1. arXiv (2026-05-21)[2026-05-25]. <https://doi.org/10.48550/arXiv.2605.22040>
- [24] Zhang L, Zhao Y Z, Yang H, et al. Enhanced fluid index modulation for integrated data and energy transfer [PP/OL]. V1. arXiv (2026-06-03)[2026-06-05]. <https://doi.org/10.48550/arXiv.2606.04537>
- [25] Ahmadzadeh M, Pakravan S, Ajib W, et al. Performance analysis of fluid antenna-assisted over-the-air federated learning under spatially correlated fading [J]. IEEE Internet of Things journal, 2026, 13(16): 37462 – 37472. DOI: 10.1109/JIOT.2026.3701371

Biographies

Huang Yinning (922104780107@njust.edu.cn) received her BEng degree from Nanjing University of Science and Technology, China in 2026. She was admitted directly to the MS program at the same university without the entrance examination in the same year. Her current research interests include fluid antenna systems, embodied intelligence, and wireless transmission optimization for mobile scenarios.

Zeng Cheng is an Associate Professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, China and the Qian-Tang Advanced Research Institute on Information, Zhejiang Key Laboratory of Intelligent Metasurface and System, China. He was a Visiting PhD Student at KTH Royal Institute of Technology, Sweden (2022 – 2023) and a Research Assistant at Chinese University of Hong Kong, Shenzhen (2024 – 2025). His current research interests include image processing and pattern recognition.

Yang Zhaohui received his bachelor's and PhD degrees from Southeast University, China in 2014 and 2018, respectively. From 2018 to 2020, he was a Postdoctoral Research Associate with the Center for Telecommunications Research, Department of Informatics, King's College London, UK. From 2020 to 2022, he was a research fellow with the Department of Electronic and Electrical Engineering, University College London, UK. He is currently a ZJU Hundred-Talent Program Research Fellow and doctoral supervisor with the Zhejiang Key Laboratory of Information Processing Communication and Networking, College of Information Science and Electronic Engineering, Zhejiang University, China. He is a recipient of the National Young Talent Program and the Young Talent Support Project of the China Association for Science and Technology. His research interests include integrated intelligent communication, sensing and computing, federated learning, fluid antenna networks, and joint resource optimization. He serves as an Associate Editor for multiple IEEE flagship journals and Guest Editor for *IEEE JSAC* and other transactions. He has received several IEEE international paper awards including the IEEE Communications Society Asia-Pacific Outstanding Young Researcher Award.



Joint Resource Allocation for Movable-Antenna-Assisted Secure Integrated Data and Energy Transfer

Xiang Pengrong¹, Li Zhiguang², Zhao Yizhe²

(1. Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Huzhou 313001, China;

2. School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

DOI: 10.12142/ZTECOM.202603005

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260828.1535.002.html>,
published online August 28, 2026

Manuscript received: 2026-07-03

Abstract: This paper investigates joint resource allocation for movable-antenna (MA)-assisted secure integrated data and energy transfer (IDET) systems with passive eavesdroppers. A field-response-based MA channel model is adopted to capture the impact of antenna positions on legitimate, energy-transfer, and eavesdropping links, while energy signals are reused as artificial noise for simultaneous wireless power transfer and secrecy enhancement. A max-min secrecy-rate problem is formulated by jointly optimizing information beamformers, energy beamformers, and MA positions under energy-harvesting, power, movable-region, and antenna-spacing constraints. To address the resulting nonconvexity, an iterative algorithm based on semidefinite relaxation, successive convex approximation, and alternating optimization is developed. Simulation results demonstrate that the proposed scheme improves the minimum secrecy rate compared with fixed-position antenna benchmarks while satisfying the energy requirements of energy receivers.

Keywords: integrated data and energy transfer; movable antenna; physical-layer security; resource allocation

Citation (Format 1): Xiang P R, Li Z G, Zhao Y Z. Joint resource allocation for movable-antenna-assisted secure integrated data and energy transfer [J]. *ZTE Communications*, 2026, 24(3): 31 – 39. DOI: 10.12142/ZTECOM.202603005

Citation (Format 2): P. R. Xiang, Z. G. Li, and Y. Z. Zhao, “Joint resource allocation for movable-antenna-assisted secure integrated data and energy transfer,” *ZTE Communications*, vol. 24, no. 3, pp. 31 – 39, Sept. 2026. doi: 10.12142/ZTECOM.202603005.

1 Introduction

Beyond-fifth-generation/sixth-generation (B5G/6G) networks are expected to support massive low-power Internet of Things (IoT) devices and ubiquitous sensing nodes, for which both reliable data transmission and continuous energy supply are essential. Integrated data and energy transfer (IDET) enables wireless information and power delivery over the same radio-frequency (RF) link, thereby extending device lifetime and reducing maintenance costs^[1–3]. However, conventional fixed-position antenna systems have limited capability to adapt the array geometry to user distribution and multipath propagation.

Security is another key challenge in IDET systems because high-power energy beams may also strengthen eavesdropping links. Physical-layer security (PLS) enhances confidentiality by exploiting channel disparities, beamforming, and artificial noise (AN)^[4–5]. In secure IDET, energy signals can be reused as AN: they are harvested by energy receivers (ERs) while in-

terfering with eavesdroppers, which enables energy transfer and secrecy protection to be jointly coordinated^[6].

Movable antenna (MA) technology provides an additional spatial-control dimension for secure IDET systems^[7–9]. By adjusting antenna positions within a finite wavelength-scale region, MAs can reconfigure multipath phase superposition, enhance legitimate links, and weaken eavesdropping links or strengthen AN leakage toward eavesdroppers^[10]. Therefore, MA position optimization can complement beamforming in improving the minimum secrecy rate.

Existing MA studies have mainly focused on channel modeling, capacity analysis, channel estimation, and position optimization^[11–13], whereas most IDET resource-allocation works rely on fixed arrays or reconfigurable intelligent surfaces^[14–15]. Although MA-assisted IDET throughput optimization was studied in Ref. [16], secure IDET with joint information transmission, energy transfer, and eavesdropping suppression remains insufficiently investigated. The resulting design is challenging because secrecy-rate constraints, position-dependent channels, and antenna-spacing constraints are all nonconvex. Recent studies have also explored practical measurement- and

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant Nos. 62571091 and 62531008.

environment-aware methods for wireless coverage enhancement. Power-measurement-based channel estimation was developed to support Intelligent Reflecting Surface (IRS) reflection design with reduced training and signaling overhead^[17]. A three-dimensional virtual-scatterer model was further proposed to reconstruct channel-gain maps from limited measurements^[18]. In addition, a Reference Signal Received Power (RSRP)-based quasi-static IRS reflection framework was investigated for low-overhead network coverage enhancement^[19]. These studies demonstrate the benefits of exploiting spatial channel information and slowly varying reconfiguration. However, they focus on IRS-assisted coverage and do not address the joint optimization of MA positions, information and energy beamforming, and secrecy performance in IDET systems.

Existing secure IDET studies such as Ref. [6] optimize information and energy beamformers for fixed array responses, whereas the MA-assisted IDET design in Ref. [16] does not consider eavesdroppers or secrecy-rate constraints. In the considered system in this paper, the continuous MA coordinates determine the data receiver (DR), ER, and eavesdropper (Eve) channel responses through multipath phase shifts and are jointly coupled with the information and energy beamformers. Consequently, the max-min secrecy-rate design involves the simultaneous handling of secrecy, energy-transfer, transmit-power, movable-region, and inter-antenna spacing constraints.

Motivated by these observations, this paper studies MA-assisted secure IDET systems and makes the following contributions.

- An MA-assisted secure IDET network model is developed, where a base station serves multiple DRs and ERs in the presence of passive eavesdroppers. A field-response channel model is used to characterize the effect of MA positions, and energy signals are reused as AN.

- A max-min secrecy-rate resource-allocation problem is formulated by jointly optimizing information beamformers, energy beamformers, and MA positions subject to energy-harvesting, transmit-power, movable-region, and minimum-distance constraints.

- An algorithm combining semidefinite relaxation (SDR), successive convex approximation (SCA), and alternating optimization (AO) is proposed, termed the SDR-SCA-AO algorithm. The beamforming subproblem is handled by semidefinite relaxation and successive convex approximation, while the MA-position subproblem is updated through iterative convex approximation of the position-dependent constraints.

- Numerical results verify the convergence of the proposed method and demonstrate the effects of transmit power, the number of MAs, ER energy thresholds, movable-region size, and minimum antenna distance on secrecy performance.

2 System Model

2.1 Movable-Antenna Channel Model

Consider the downlink MA-assisted secure IDET network

shown in Fig. 1. The base station is equipped with M movable antennas. The position of the m -th antenna is denoted by $\mathbf{r}_m = [x_m, y_m]^T$, which is confined to a two-dimensional region $\mathcal{C} = [-A/2, A/2] \times [-A/2, A/2]$, where A denotes the side length of the movable region. The entire antenna position set is represented as $\tilde{\mathbf{r}} = \{\mathbf{r}_m\}_{m=1}^M$. To avoid antenna collision and strong mutual coupling, the antenna positions are subject to the following constraints:

$$\begin{aligned} \mathbf{r}_m &\in \mathcal{C}, \quad \forall m, \\ \|\mathbf{r}_m - \mathbf{r}_i\|_2 &\geq D, \quad m \neq i \end{aligned} \quad (1)$$

where D is the minimum distance between any two antennas.

For any receiving node u , including a DR, an ER, or an Eve, let L_u denote the number of propagation paths. Since this paper optimizes the transmit-side MA positions at the base station, all angular parameters are defined at the transmit-array side. Specifically, $\theta_{u,l}$ denotes the azimuth angle of departure of the l -th path from the base station to node u , while $\phi_{u,l}$ denotes the corresponding elevation angle of departure. Because each receiving node is equipped with a single antenna, no receive-array response is considered. The phase difference of the l -th path relative to the reference point of the base station is written as

$$\rho_{u,l}(\mathbf{r}_m) = \frac{2\pi}{\lambda} (x_m \sin\theta_{u,l} \cos\phi_{u,l} + y_m \cos\phi_{u,l}) \quad (2),$$

where λ denotes the carrier wavelength. The field-response vector at the m -th MA is given by

$$\mathbf{f}_u(\mathbf{r}_m) = [e^{j\rho_{u,1}(\mathbf{r}_m)}, \dots, e^{j\rho_{u,L_u}(\mathbf{r}_m)}]^T \quad (3).$$

By stacking the field-response vectors of all M antennas, we obtain:

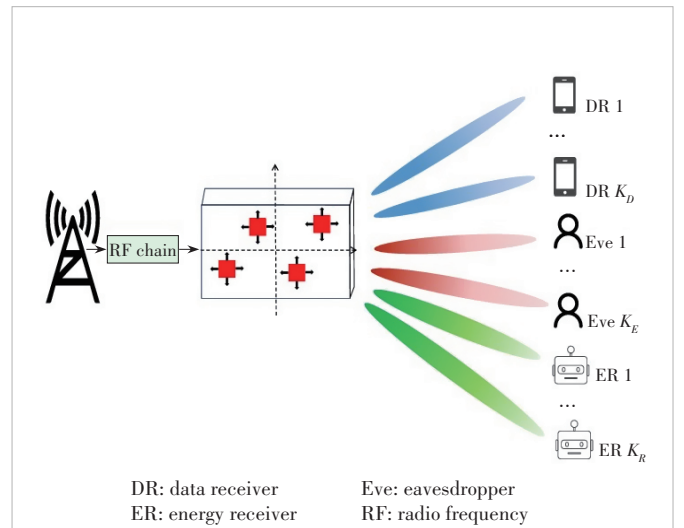


Figure 1. Illustration of an MA-assisted secure IDET network

$$\mathbf{F}_u(\tilde{\mathbf{r}}) = [\mathbf{f}_u(\mathbf{r}_1), \dots, \mathbf{f}_u(\mathbf{r}_M)] \in \mathbb{C}^{L_u \times M} \quad (4).$$

Let $\mathbf{a}_u \in \mathbb{C}^{L_u \times 1}$ denote the complex path-gain vector of node u . The equivalent channel vector from the base station to node u is then expressed as

$$\begin{aligned} \mathbf{h}_u(\tilde{\mathbf{r}}) &= \mathbf{F}_u^H(\tilde{\mathbf{r}}) \mathbf{a}_u, \\ a_{u,l} &\sim \mathcal{CN}\left(0, \frac{\beta_0 d_u^{-\alpha}}{L_u}\right), \quad l = 1, \dots, L_u \end{aligned} \quad (5),$$

where d_u denotes the propagation distance, α is the path-loss exponent, and β_0 is the path loss at the reference distance. It follows from Eq. (5) that the MA positions affect the channel through exponential phase terms, and position changes on the wavelength scale can alter the coherent superposition of multipath signals.

The channels are assumed to remain quasi-static within each transmission block. The BS acquires the CSI of the DRs, ERs, and Eves through pilot-assisted channel estimation. Under time-division duplexing, mutually orthogonal uplink pilot sequences are assigned to the terminals, and the corresponding downlink CSI is obtained using channel reciprocity after radio-frequency-chain calibration. The resulting channel estimates are treated as perfectly known during each resource-allocation interval.

2.2 Secure IDET Signal Model

Consider the DR set $\mathcal{K}_D = \{1, \dots, K_D\}$, the ER set $\mathcal{K}_R = \{1, \dots, K_R\}$, and the Eve set $\mathcal{K}_E = \{1, \dots, K_E\}$. For simplicity, $\mathbf{h}_{k_D}$, $\mathbf{g}_{k_R}$, and $\mathbf{q}_{k_E}$ denote the channel vectors from the base station to the k_D -th DR, the k_R -th ER, and the k_E -th Eve, respectively. They are all given by Eq. (5) and depend on $\tilde{\mathbf{r}}$. The base station simultaneously transmits information streams and energy streams over the same time-frequency resource:

$$\mathbf{x} = \sum_{k_D \in \mathcal{K}_D} \mathbf{w}_{k_D} s_{k_D}^D + \sum_{k_R \in \mathcal{K}_R} \mathbf{v}_{k_R} s_{k_R}^R \quad (6),$$

where $\mathbf{w}_{k_D}$ is the information beamformer, $\mathbf{v}_{k_R}$ is the energy beamformer, and $s_{k_D}^D$ and $s_{k_R}^R$ have unit power and are mutually independent. The energy streams act as power-carrying signals at ERs, as AN at Eves, and as controllable interference at DRs.

The received signal-to-interference-plus-noise ratio (SINR) at the k_D -th DR is

$$\gamma_{k_D}^D = \frac{|\mathbf{h}_{k_D}^H \mathbf{w}_{k_D}|^2}{\sum_{i_D \neq k_D} |\mathbf{h}_{k_D}^H \mathbf{w}_{i_D}|^2 + \sum_{i_R} |\mathbf{h}_{k_D}^H \mathbf{v}_{i_R}|^2 + \sigma_{k_D}^2} \quad (7).$$

The SINR at the k_E -th Eve when it eavesdrops on the information stream intended for DR k_D is

$$\gamma_{k_D, k_E}^E = \frac{|\mathbf{q}_{k_E}^H \mathbf{w}_{k_D}|^2}{\sum_{i_D \neq k_D} |\mathbf{q}_{k_E}^H \mathbf{w}_{i_D}|^2 + \sum_{i_R} |\mathbf{q}_{k_E}^H \mathbf{v}_{i_R}|^2 + \sigma_{k_E}^2} \quad (8).$$

The corresponding legitimate and eavesdropping rates are

$$\begin{aligned} R_{k_D}^D &= \log_2(1 + \gamma_{k_D}^D), \\ R_{k_D, k_E}^E &= \log_2(1 + \gamma_{k_D, k_E}^E) \end{aligned} \quad (9).$$

Considering the worst-case Eve, the secrecy rate of the k_D -th DR is defined as

$$R_{k_D}^{\text{sec}} = \max \left\{ R_{k_D}^D - \max_{k_E \in \mathcal{K}_E} R_{k_D, k_E}^E, 0 \right\} \quad (10),$$

where the maximum with zero ensures a nonnegative secrecy rate.

2.3 IDET Receiver Structure and Energy Harvesting Model

A separate IDET receiver architecture is adopted, in which DRs perform only information decoding and ERs perform only energy harvesting. For any ER $j \in \mathcal{K}_R$, let $P_{\text{in},j}$ denote the total RF power received at the input of its rectifier. This power is contributed by both information beams and energy beams. From Eq. (6), the RF input power of the k_R -th ER can be written as

$$P_{\text{in},k_R} = \sum_{i_D \in \mathcal{K}_D} |\mathbf{g}_{k_R}^H \mathbf{w}_{i_D}|^2 + \sum_{i_R \in \mathcal{K}_R} |\mathbf{g}_{k_R}^H \mathbf{v}_{i_R}|^2 \quad (11).$$

A practical rectifier circuit exhibits a saturation effect. Its energy output can be described by the nonlinear model

$$\Psi_j(P_{\text{in},j}) = \frac{\Omega_j}{1 + \exp[-a_j(P_{\text{in},j} - b_j)]} - \frac{\Omega_j}{1 + \exp(a_j b_j)} \quad (12),$$

where Ω_j is the saturation power, and a_j and b_j are determined by the rectifier circuit. Since $\Psi_j(\cdot)$ is monotonically increasing with respect to $P_{\text{in},j}$, the constraint $\Psi_j(P_{\text{in},j}) \geq E_{\text{th},j}$ can be equivalently rewritten as

$$P_{\text{in},j} \geq P_{\text{th},j}, \quad j \in \mathcal{K}_R \quad (13),$$

where $P_{\text{th},j}$ is the equivalent radio-frequency input-power threshold obtained by inverting $E_{\text{th},j}$. In the subsequent optimization, Eq. (13) is used to represent the energy-harvesting constraint, thereby reducing computational complexity while preserving the physical meaning of nonlinear energy harvesting.

3 Joint Resource Allocation for Secure IDET

3.1 Problem Formulation

The objective is to maximize the minimum secrecy rate among all DRs while satisfying the ER energy constraints, the total transmit-power constraint, and the physical constraints of MAs. The original problem is formulated as

$$\begin{aligned}
(\text{P0}): \quad & \max_{\{\mathbf{w}_{k_D}, \{\mathbf{v}_{k_R}\}, \tilde{\mathbf{r}}, k_D \in \mathcal{K}_D} \min R_{k_D}^{\text{sec}} \\
\text{s.t.} \quad & P_{\text{in}, k_R} \geq P_{\text{th}, k_R}, \quad \forall k_R \in \mathcal{K}_R, \\
& \sum_{k_D} \|\mathbf{w}_{k_D}\|^2 + \sum_{k_R} \|\mathbf{v}_{k_R}\|^2 \leq P, \\
& \mathbf{r}_m \in \mathcal{C}, \quad \forall m, \\
& \|\mathbf{r}_m - \mathbf{r}_i\|_2 \geq D, \quad \forall m \neq i
\end{aligned} \quad (14).$$

By introducing an auxiliary variable τ and expanding the worst-case eavesdropping constraint, (P0) can be reformulated as

$$\begin{aligned}
(\text{P1}): \quad & \max_{\{\mathbf{w}_{k_D}, \{\mathbf{v}_{k_R}\}, \tilde{\mathbf{r}}, \tau} \tau \\
\text{s.t.} \quad & R_{k_D}^D - R_{k_D, k_E}^E \geq \tau, \quad \forall k_D, k_E, \\
& P_{\text{in}, k_R} \geq P_{\text{th}, k_R}, \quad \forall k_R, \\
& \sum_{k_D} \|\mathbf{w}_{k_D}\|^2 + \sum_{k_R} \|\mathbf{v}_{k_R}\|^2 \leq P, \\
& \mathbf{r}_m \in \mathcal{C}, \quad \forall m, \\
& \|\mathbf{r}_m - \mathbf{r}_i\|_2 \geq D, \quad \forall m \neq i
\end{aligned} \quad (15).$$

Problem (P1) is nonconvex for the following reasons. First, the secrecy-rate constraint contains a difference between two logarithmic rates, and each rate further includes a fractional SINR. Second, the beamforming variables and MA-position variables are coupled through the channel vectors. Third, the minimum antenna-distance constraint $\|\mathbf{r}_m - \mathbf{r}_i\|_2 \geq D$ defines a nonconvex feasible set. Therefore, an AO framework is adopted, in which the variables are divided into the beamforming-matrix block and the MA-position block, and SDR and SCA are used to handle the corresponding subproblems.

3.2 Beamforming Optimization with Fixed MA Positions

Given $\tilde{\mathbf{r}}$, all channel vectors are fixed. First, define the information and energy beamforming matrices as

$$\begin{aligned}
\mathbf{W}_{k_D} &= \mathbf{w}_{k_D} \mathbf{w}_{k_D}^H, \quad \mathbf{W}_{k_D} \geq 0, \text{rank}(\mathbf{W}_{k_D}) = 1, \\
\mathbf{V}_{k_R} &= \mathbf{v}_{k_R} \mathbf{v}_{k_R}^H, \quad \mathbf{V}_{k_R} \geq 0, \text{rank}(\mathbf{V}_{k_R}) = 1
\end{aligned} \quad (16).$$

Define the channel matrices as

$$\begin{aligned}
\mathbf{H}_{k_D} &= \mathbf{h}_{k_D} \mathbf{h}_{k_D}^H, \\
\mathbf{G}_{k_R} &= \mathbf{g}_{k_R} \mathbf{g}_{k_R}^H, \\
\mathbf{Q}_{k_E} &= \mathbf{q}_{k_E} \mathbf{q}_{k_E}^H
\end{aligned} \quad (17).$$

Then $|\mathbf{h}^H \mathbf{w}|^2 = \text{Tr}(\mathbf{h} \mathbf{h}^H \mathbf{W})$, and the corresponding beamforming power is $\text{Tr}(\mathbf{W})$. The legitimate-link rate can be written in the following difference-of-concave form:

$$R_{k_D}^D = \log_2 A_{k_D}(\mathbf{W}, \mathbf{V}) - \log_2 B_{k_D}(\mathbf{W}, \mathbf{V}) \quad (18),$$

where

$$\begin{aligned}
A_{k_D} &= \sum_{i_D} \text{Tr}(\mathbf{H}_{k_D} \mathbf{W}_{i_D}) + \sum_{i_R} \text{Tr}(\mathbf{H}_{k_D} \mathbf{V}_{i_R}) + \sigma_{k_D}^2, \\
B_{k_D} &= \sum_{i_D \neq k_D} \text{Tr}(\mathbf{H}_{k_D} \mathbf{W}_{i_D}) + \sum_{i_R} \text{Tr}(\mathbf{H}_{k_D} \mathbf{V}_{i_R}) + \sigma_{k_D}^2
\end{aligned} \quad (19).$$

Similarly, the Eve rate can be written as

$$R_{k_D, k_E}^E = \log_2 C_{k_D, k_E}(\mathbf{W}, \mathbf{V}) - \log_2 D_{k_D, k_E}(\mathbf{W}, \mathbf{V}) \quad (20),$$

where

$$\begin{aligned}
C_{k_D, k_E} &= \sum_{i_D} \text{Tr}(\mathbf{Q}_{k_E} \mathbf{W}_{i_D}) + \sum_{i_R} \text{Tr}(\mathbf{Q}_{k_E} \mathbf{V}_{i_R}) + \sigma_{k_E}^2, \\
D_{k_D, k_E} &= \sum_{i_D \neq k_D} \text{Tr}(\mathbf{Q}_{k_E} \mathbf{W}_{i_D}) + \sum_{i_R} \text{Tr}(\mathbf{Q}_{k_E} \mathbf{V}_{i_R}) + \sigma_{k_E}^2
\end{aligned} \quad (21).$$

Thus, the secrecy-rate difference of any DR-Eve pair is

$$\begin{aligned}
R_{k_D}^D - R_{k_D, k_E}^E &= \log_2 A_{k_D} - \log_2 B_{k_D} - \\
& \log_2 C_{k_D, k_E} + \log_2 D_{k_D, k_E}
\end{aligned} \quad (22).$$

In Eq. (22), $\log_2 A_{k_D}$ and $\log_2 D_{k_D, k_E}$ are concave functions, whereas $-\log_2 B_{k_D}$ and $-\log_2 C_{k_D, k_E}$ are convex functions. This structure leads to nonconvex constraints. Since $\log_2(x)$ is concave, its first-order upper bound at the n -th SCA iteration point is

$$\log_2 x \leq \log_2 x^{(n)} + \frac{x - x^{(n)}}{x^{(n)} \ln 2}, \quad x > 0 \quad (23).$$

By replacing $\log_2 B_{k_D}$ and $\log_2 C_{k_D, k_E}$ with their upper bounds, a concave lower bound of the secrecy-rate difference can be obtained as

$$\begin{aligned}
\hat{R}_{k_D, k_E}^{(n)} &= \log_2 A_{k_D} + \log_2 D_{k_D, k_E} - \\
& \left[\log_2 B_{k_D}^{(n)} + \frac{B_{k_D} - B_{k_D}^{(n)}}{B_{k_D}^{(n)} \ln 2} \right] - \\
& \left[\log_2 C_{k_D, k_E}^{(n)} + \frac{C_{k_D, k_E} - C_{k_D, k_E}^{(n)}}{C_{k_D, k_E}^{(n)} \ln 2} \right]
\end{aligned} \quad (24).$$

Here, $B_{k_D}^{(n)}$ and $C_{k_D, k_E}^{(n)}$ are computed using the solution of the previous SCA iteration. Since $\hat{R}_{k_D, k_E}^{(n)} \leq R_{k_D}^D - R_{k_D, k_E}^E$ and equality holds at the current point, $\hat{R}_{k_D, k_E}^{(n)} \geq \tau$ is used to guarantee conservative feasibility.

The energy and power constraints can be expressed in terms of the matrix variables as

$$\begin{aligned}
\sum_{i_D} \text{Tr}(\mathbf{G}_{k_R} \mathbf{W}_{i_D}) + \sum_{i_R} \text{Tr}(\mathbf{G}_{k_R} \mathbf{V}_{i_R}) &\geq P_{\text{th}, k_R}, \quad \forall k_R, \\
\sum_{k_D} \text{Tr}(\mathbf{W}_{k_D}) + \sum_{k_R} \text{Tr}(\mathbf{V}_{k_R}) &\leq P
\end{aligned} \quad (25).$$

After removing the nonconvex rank-one constraints, the beamforming subproblem in the n -th SCA iteration becomes

$$\begin{aligned}
 (\text{P2}): \quad & \max_{\{\mathbf{W}_{k_D}, \mathbf{V}_{k_E}, \tau\}} \tau \\
 \text{s.t.} \quad & \hat{R}_{k_D, k_E}^{(n)} \geq \tau, \quad \forall k_D, k_E, \\
 & \mathbf{W}_{k_D} \geq 0, \quad \forall k_D, \\
 & \mathbf{V}_{k_E} \geq 0, \quad \forall k_E.
 \end{aligned} \quad (26)$$

The objective function is linear, the constraint $\hat{R}_{k_D, k_E}^{(n)} \geq \tau$ is a superlevel-set constraint of a concave function, and the remaining constraints are linear matrix inequalities or linear constraints. Hence, (P2) is a convex semidefinite programming (SDP) problem. If the obtained $\mathbf{W}_{k_D}$ and $\mathbf{V}_{k_E}$ are all rank-one matrices, the beamformers can be recovered directly through the principal eigenvectors. If high-rank solutions are obtained, Gaussian randomization is used to generate candidate beamformers, and the candidate satisfying the power and energy constraints while achieving the highest secrecy rate is selected.

For each high-rank SDR matrix, eigenvalue decomposition is first performed. In each trial, an independent complex Gaussian vector is generated and mapped through the eigenvector matrix and the square root of the eigenvalue matrix to construct candidate information and energy beamformers. The candidates are then power-scaled and evaluated under the original constraints. Among N_{GR} trials, the feasible candidate set yielding the largest original minimum secrecy rate is selected^[20–21].

3.3 MA Position Optimization with Fixed Beamformers

Given the information and energy beamformers, the MA positions are optimized to reconfigure the legitimate, energy-transfer, and eavesdropping links. Define the useful signal power and the interference-plus-noise power of any node u with respect to the target information stream k_D as

$$\begin{aligned}
 S_{u, k_D}(\tilde{\mathbf{r}}) &= \left| \mathbf{h}_u^H(\tilde{\mathbf{r}}) \mathbf{w}_{k_D} \right|^2, \\
 I_{u, k_D}(\tilde{\mathbf{r}}) &= \sum_{i_D \neq k_D} \left| \mathbf{h}_u^H(\tilde{\mathbf{r}}) \mathbf{w}_{i_D} \right|^2 + \\
 &\quad \sum_{i_R} \left| \mathbf{h}_u^H(\tilde{\mathbf{r}}) \mathbf{v}_{i_R} \right|^2 + \sigma_u^2
 \end{aligned} \quad (27)$$

When u is a DR, $\mathbf{h}_u$ denotes a legitimate channel; when u is an Eve, $\mathbf{h}_u$ denotes an eavesdropping channel. Let

$$\begin{aligned}
 A_{u, k_D}(\tilde{\mathbf{r}}) &= S_{u, k_D}(\tilde{\mathbf{r}}) + I_{u, k_D}(\tilde{\mathbf{r}}), \\
 B_{u, k_D}(\tilde{\mathbf{r}}) &= I_{u, k_D}(\tilde{\mathbf{r}})
 \end{aligned} \quad (28)$$

The unified rate expression is therefore

$$R_{u, k_D}(\tilde{\mathbf{r}}) = \log_2 A_{u, k_D}(\tilde{\mathbf{r}}) - \log_2 B_{u, k_D}(\tilde{\mathbf{r}}) \quad (29)$$

The secrecy-rate difference between the k_D -th DR and the

k_E -th Eve is

$$\begin{aligned}
 R_{k_D, k_E}^{\text{sec}}(\tilde{\mathbf{r}}) &= \log_2 A_{k_D, k_D}(\tilde{\mathbf{r}}) - \log_2 B_{k_D, k_D}(\tilde{\mathbf{r}}) - \\
 &\quad \log_2 A_{k_E, k_D}(\tilde{\mathbf{r}}) + \log_2 B_{k_E, k_D}(\tilde{\mathbf{r}})
 \end{aligned} \quad (30)$$

According to Eq. (23), at the current iteration point $\tilde{\mathbf{r}}^{(n)}$, we have

$$\begin{aligned}
 \log_2 B_{k_D, k_D}(\tilde{\mathbf{r}}) &\leq \log_2 B_{k_D, k_D}^{(n)} + \frac{B_{k_D, k_D}(\tilde{\mathbf{r}}) - B_{k_D, k_D}^{(n)}}{B_{k_D, k_D}^{(n)} \ln 2}, \\
 \log_2 A_{k_E, k_D}(\tilde{\mathbf{r}}) &\leq \log_2 A_{k_E, k_D}^{(n)} + \frac{A_{k_E, k_D}(\tilde{\mathbf{r}}) - A_{k_E, k_D}^{(n)}}{A_{k_E, k_D}^{(n)} \ln 2}
 \end{aligned} \quad (31)$$

where $B_{k_D, k_D}^{(n)} = B_{k_D, k_D}(\tilde{\mathbf{r}}^{(n)})$ and $A_{k_E, k_D}^{(n)} = A_{k_E, k_D}(\tilde{\mathbf{r}}^{(n)})$. Hence, a conservative lower bound of the secrecy-rate difference is

$$\begin{aligned}
 \hat{R}_{k_D, k_E}^{\text{pos}, (n)} &= \log_2 A_{k_D, k_D}(\tilde{\mathbf{r}}) + \log_2 B_{k_E, k_D}(\tilde{\mathbf{r}}) - \\
 &\quad \left[\log_2 B_{k_D, k_D}^{(n)} + \frac{B_{k_D, k_D}(\tilde{\mathbf{r}}) - B_{k_D, k_D}^{(n)}}{B_{k_D, k_D}^{(n)} \ln 2} \right] - \\
 &\quad \left[\log_2 A_{k_E, k_D}^{(n)} + \frac{A_{k_E, k_D}(\tilde{\mathbf{r}}) - A_{k_E, k_D}^{(n)}}{A_{k_E, k_D}^{(n)} \ln 2} \right]
 \end{aligned} \quad (32)$$

The radio-frequency input power of the k_R -th ER can be written as

$$\Gamma_{k_R}(\tilde{\mathbf{r}}) = \sum_{i_D} \left| \mathbf{g}_{k_R}^H(\tilde{\mathbf{r}}) \mathbf{w}_{i_D} \right|^2 + \sum_{i_R} \left| \mathbf{g}_{k_R}^H(\tilde{\mathbf{r}}) \mathbf{v}_{i_R} \right|^2 \quad (33)$$

Let $\mathbf{z} = [x_1, y_1, \dots, x_M, y_M]^T$ denote the position vector. For any beamformer $\mathbf{b}$, define the power function as

$$p_{u, \mathbf{b}}(\mathbf{z}) = \left| \mathbf{h}_u^H(\mathbf{z}) \mathbf{b} \right|^2 \quad (34)$$

Its gradient with respect to any coordinate $\chi \in \{x_m, y_m\}$ is

$$\frac{\partial p_{u, \mathbf{b}}}{\partial \chi} = 2 \text{Re} \left(\overline{\mathbf{h}_u^H \mathbf{b}} \frac{\partial (\mathbf{h}_u^H \mathbf{b})}{\partial \chi} \right) \quad (35)$$

When $\chi = x_m$, it follows from Eqs. (2)–(5) that

$$\frac{\partial (\mathbf{h}_u^H \mathbf{b})}{\partial x_m} = j \frac{2\pi}{\lambda} b_m \sum_{l=1}^{L_u} a_{u, l}^* e^{j \frac{2\pi}{\lambda} \rho_{u, l}(r_m)} \times \sin \theta_{u, l} \cos \phi_{u, l} \quad (36)$$

where $\sin \theta_{u, l} \cos \phi_{u, l}$ is replaced with $\cos \phi_{u, l}$ when $\chi = y_m$.

Therefore, the first-order approximation of the ER input power at the n -th iteration point is

$$\begin{aligned}
 \hat{\Gamma}_{k_R}^{(n)}(\mathbf{z}) &= \Gamma_{k_R}(\mathbf{z}^{(n)}) + \nabla \Gamma_{k_R}(\mathbf{z}^{(n)})^T (\mathbf{z} - \mathbf{z}^{(n)}), \\
 \nabla \Gamma_{k_R}(\mathbf{z}^{(n)}) &= \sum_{i_D} \nabla p_{k_R, \mathbf{w}_{i_D}}(\mathbf{z}^{(n)}) + \sum_{i_R} \nabla p_{k_R, \mathbf{v}_{i_R}}(\mathbf{z}^{(n)})
 \end{aligned} \quad (37)$$

Thus, the energy constraint in the current SCA iteration becomes

$$\hat{I}_{k_R}^{(n)}(z) \geq P_{th, k_R}, \quad \forall k_R \quad (38).$$

For the minimum antenna distance, since $\|\mathbf{r}_m - \mathbf{r}_i\|_2^2$ is a convex function, its first-order lower bound is

$$\begin{aligned} & \|\mathbf{r}_m^{(n)} - \mathbf{r}_i^{(n)}\|_2^2 + 2(\mathbf{r}_m^{(n)} - \mathbf{r}_i^{(n)})^T \times \\ & \left[(\mathbf{r}_m - \mathbf{r}_i) - (\mathbf{r}_m^{(n)} - \mathbf{r}_i^{(n)}) \right] \geq D^2, \quad \forall m \neq i \end{aligned} \quad (39).$$

Accordingly, the approximate MA-position subproblem with fixed beamformers is formulated as

$$\begin{aligned} & \text{(P3): } \max_{z, \tau} \tau \\ & \text{s.t. } \hat{R}_{k_D, k_E}^{\text{pos}, (n)} \geq \tau, \quad \forall k_D, k_E, \\ & \hat{I}_{k_R}^{(n)}(z) \geq P_{th, k_R}, \quad \forall k_R \\ & \mathbf{r}_m \in \mathcal{C}, \quad \forall m \end{aligned} \quad (40).$$

Thus, (P3) is the SCA approximation problem at the current iteration point and can be solved by standard convex optimization tools. After an optimal solution is obtained, it is used as the next SCA linearization point, and the iterations continue until the objective value converges.

3.4 Overall Algorithm and Complexity Analysis

Based on the two subproblems above, an SDR-SCA-AO joint resource-allocation algorithm is proposed for secure IDET, as shown in Algorithm 1. The algorithm first initializes MA positions that satisfy the region and distance constraints and constructs feasible beamformers that satisfy the power constraint. In each outer iteration, the beamforming SDP subproblem is solved with fixed MA positions, and the position SCA subproblem is then solved with fixed beamformers. The iterations terminate when the change in the minimum secrecy rate is below a preset threshold.

Algorithm 1. SDR-SCA-AO joint resource allocation algorithm for secure IDET

Input: Initial MA positions, initial information and energy beamformers, convergence threshold ϵ , and maximum number of iterations $N_{\max}$. The number of Gaussian-randomization trials is denoted by N_{GR} .

Output: Optimized information beamformers, energy beamformers, MA positions, and objective value τ^* .

1. Initialize the iteration index $n = 0$ and compute the initial objective value $\tau^{(0)}$.
2. While $n < N_{\max}$:
 - Fix the MA positions $\tilde{\mathbf{r}}^{(n)}$.
 - Construct the corresponding channel matrices $\mathbf{H}_{k_D}$, $\mathbf{G}_{k_R}$, and $\mathbf{Q}_{k_E}$.
 - Solve the beamforming optimization subproblem using SDR and SCA, and obtain $\{\mathbf{W}_{k_D}^{(n+1)}\}$, $\{\mathbf{V}_{k_R}^{(n+1)}\}$, and the

objective value $\tau_{\text{beam}}^{(n+1)}$.

- If all $\mathbf{W}_{k_D}^{(n+1)}$ and $\mathbf{V}_{k_R}^{(n+1)}$ are rank-one matrices, recover the information beamformers $\{\mathbf{w}_{k_D}^{(n+1)}\}$ and energy beamformers $\{\mathbf{v}_{k_R}^{(n+1)}\}$ through eigenvalue decomposition; otherwise, construct an approximately feasible beamforming solution using Gaussian randomization.
- For a high-rank solution, perform N_{GR} Gaussian-randomization trials, scale the candidate beamformers, discard infeasible candidates, and retain the feasible candidate with the highest original minimum secrecy rate.
- Fix the updated information beamformers $\{\mathbf{w}_{k_D}^{(n+1)}\}$ and energy beamformers $\{\mathbf{v}_{k_R}^{(n+1)}\}$.
- Solve the antenna-positioning subproblem using SCA, and obtain the updated MA positions $\tilde{\mathbf{r}}^{(n+1)}$ and the objective value $\tau^{(n+1)}$.
- Evaluate the candidate solution using the original objective and all original constraints. Accept the update only if it is feasible and its minimum secrecy rate is no lower than that of the previous accepted iterate; otherwise, retain the previous feasible solution.
- If $|\tau^{(n+1)} - \tau^{(n)}| \leq \epsilon$, terminate the loop.
- Set $n \leftarrow n + 1$.

3. Output the final solution.

The objective-value sequence generated by Algorithm 1 is monotonically nondecreasing. Let $\tau^{(n)}$ denote the objective value at the n -th accepted iterate. The SCA surrogate functions in (P2) and (P3) are tight at the current point, and the current solution remains feasible for the corresponding approximated problem. Solving the two subproblems and applying the acceptance criterion specified in Algorithm 1 thus yields $\tau^{(n+1)} \geq \tau^{(n)}$. Moreover, under the finite total transmit-power constraint, the achievable secrecy rate is upper bounded. Hence, $\{\tau^{(n)}\}$ is a bounded nondecreasing sequence, and Algorithm 1 converges.

Let I_{AO} denote the number of outer AO iterations, and let I_{B} and I_{P} denote the numbers of inner iterations for the beamforming SCA and position SCA, respectively. The beamforming subproblem contains $(K_D + K_R)$ positive semidefinite matrices of size $M \times M$, and its dominant complexity can be approximately expressed as

$$C_{\text{B}} = \mathcal{O}\left(\left((K_D + K_R)M^2\right)^3\right) \quad (41).$$

The dimension of the position subproblem is approximately $2M$, and it contains $K_D K_E$ secrecy constraints, K_R energy constraints, and $M(M-1)/2$ distance constraints. Its complexity can be roughly characterized as

$$C_{\text{P}} = \mathcal{O}\left((2M)^3\right) = \mathcal{O}(M^3) \quad (42).$$

Therefore, the total complexity is

$$\mathcal{O}\{I_{AO}(I_B C_B + I_P C_P)\} \quad (43).$$

It can be observed from Eq. (43) that the main computational burden lies in the beamforming SDP subproblem.

4 Simulation Results

To verify the effectiveness of the proposed algorithm, we consider a downlink MA-assisted secure IDET network. The base station is located at the origin and is equipped with $M = 6$ MAs. The system contains $K_D = 2$ DRs, $K_R = 2$ ERs, and $K_E = 2$ Eves. The carrier wavelength is set to $\lambda = 0.1$ m, the side length of the MA movable region is $A = 3\lambda$, and the minimum antenna distance is $D = \lambda/2$. The path-loss exponent is $\alpha = 2.4$, the reference-distance path loss is -40 dB, the noise power is -80 dBm, and the maximum transmit power of the base station is set to 20 dBm unless otherwise specified. The equivalent energy threshold of each ER is set to 0.01 mW by default. The simulation results are averaged over multiple independent channel realizations. The main parameters are listed in Table 1.

The proposed algorithm is referred to as the MA-PLS scheme. Three benchmarks are considered for comparison.

1) Fixed-position antenna (FPA) scheme: the BS adopts a uniform planar array with fixed antenna positions, and only the information and energy beamformers are optimized.

2) Without-eavesdropper (No-Eve) scheme: Eves are absent and the PLS constraints are removed, which provides an ideal reference without security threats.

3) FPA-No-Eve scheme: fixed antenna positions and the No-Eve setting are both adopted, representing a conventional IDET baseline.

These benchmarks isolate the effects of MA position optimization, PLS constraints, and energy-signal reuse as AN.

Table 1. Main simulation parameters

Parameter	Meaning	Value
M	Number of MAs	6
K_D, K_R	Numbers of DRs and ERs	2, 2
K_E	Number of Eves	2
A	Side length of movable region	3λ
D	Minimum antenna distance	$\lambda/2$
P	Transmit power	20 dBm
σ^2	Noise power	-80 dBm
α	Path-loss exponent	2.4
P_{th}	Equivalent energy threshold	0.01 mW
λ	Carrier wavelength	0.1 m
L_u	Number of propagation paths	10
N_{CR}	Gaussian-randomization trials	1 000

DR: data receiver ER: energy receiver MA: movable antenna

Fig. 2 shows the convergence behavior of the proposed algorithm. The minimum secrecy rate increases rapidly in the first few iterations and then stabilizes, indicating that the SDR-SCA-AO framework effectively coordinates beamforming and MA-position updates. A larger number of MAs yields a higher converged value because more spatial degrees of freedom are available. This behavior occurs because the early iterations substantially improve the initial beamformers and antenna geometry, whereas the gain produced by each subsequent update becomes smaller near convergence.

Fig. 3 shows the minimum secrecy rate versus the transmit power, where P_{max} varies from 10 dBm to 30 dBm in steps of 2 dB. The proposed MA-PLS scheme consistently outperforms the FPA scheme because MA position optimization provides additional spatial degrees of freedom for enhancing legitimate links and suppressing eavesdropping links. The additional

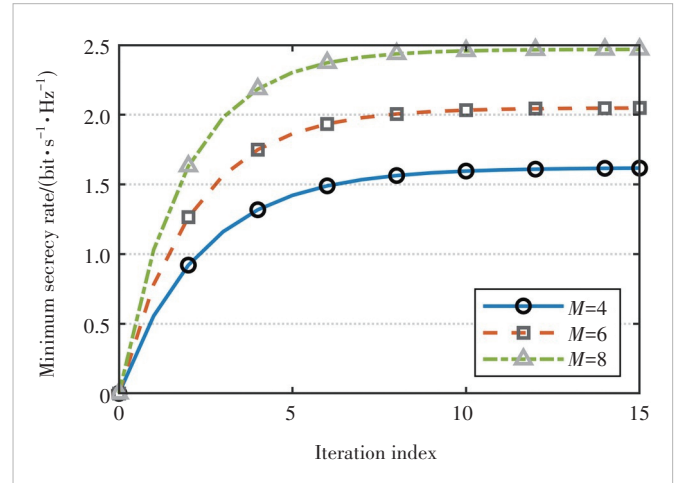


Figure 2. Convergence analysis of the SDR-SCA-AO algorithm

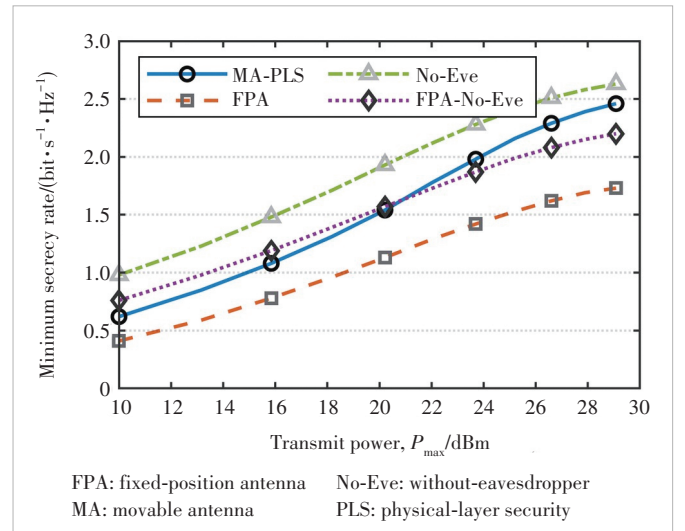


Figure 3. Minimum secrecy rate versus base-station transmit power under different schemes

transmit power benefits both desired transmission and energy-signal-based artificial noise; the MA scheme can use its reconfigurable geometry to direct these gains toward the DRs and interference toward the Eves more effectively than the fixed array.

Fig. 4 investigates the impact of the number of MAs on the minimum secrecy rate. Increasing the number of antennas improves both beamforming flexibility and spatial reconfigurability, leading to a higher minimum secrecy rate. The proposed MA-PLS scheme maintains clear gains over fixed-position benchmarks. Each additional MA contributes not only an extra beamforming coefficient but also an optimizable spatial coordinate, which explains the increasing advantage over the fixed-position benchmarks.

Fig. 5 evaluates the effect of the ER energy threshold on system performance. A higher threshold requires more power and spatial resources for energy transfer, leaving fewer resources for information transmission and secrecy enhancement. Consequently, the minimum secrecy rate decreases, while the proposed scheme remains superior to the FPA benchmark. Although the energy signals also act as artificial noise, a stringent ER requirement reduces the remaining beamforming freedom; MA positioning mitigates this loss by strengthening the ER channels and lowering the power cost of satisfying the threshold.

Fig. 6 illustrates the influence of the normalized movable-region side length A/λ on the secrecy rate. Enlarging the movable region provides more feasible antenna positions, which improves channel reconfiguration and secrecy performance. The gain gradually saturates because of the finite transmit power and fixed network topology. A larger region allows the MAs to explore more distinct multipath phase combinations, whereas the saturation indicates that the performance eventually becomes limited by the finite power budget and the available propagation paths.

Fig. 7 shows the impact of the normalized minimum antenna distance D/λ on the achievable secrecy rate. A larger spacing requirement reduces the feasible position region and weakens the flexibility of array-geometry optimization, thereby decreasing the minimum secrecy rate. Even under this constraint, the proposed scheme outperforms the FPA benchmark. The larger spacing removes some channel-favorable antenna configurations from the feasible set, but the proposed scheme can still optimize the remaining positions jointly with the beamformers.

5 Conclusions

This paper investigated joint resource allocation for MA-assisted secure IDET systems. By exploiting a field-response MA channel model, energy signals were reused as AN, and a max-min secrecy-rate problem was formulated to jointly optimize information beamformers, energy beamformers, and MA positions under practical energy-harvesting and antenna-position constraints. An SDR-SCA-AO algorithm was developed to solve the resulting nonconvex problem. Simulation re-

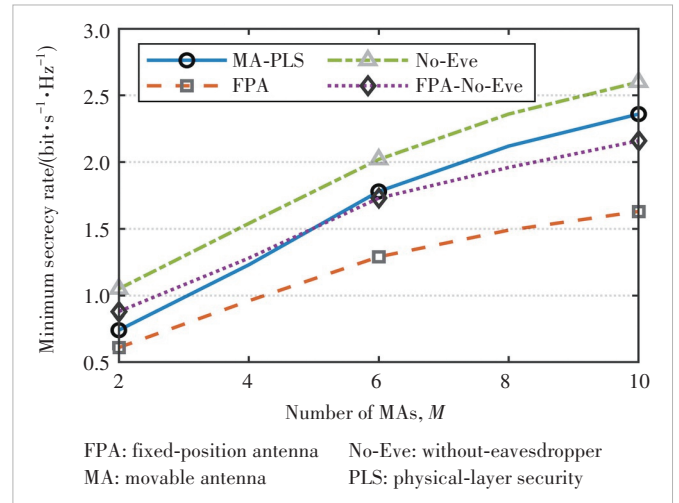


Figure 4. Minimum secrecy rate versus the number of MAs under different schemes

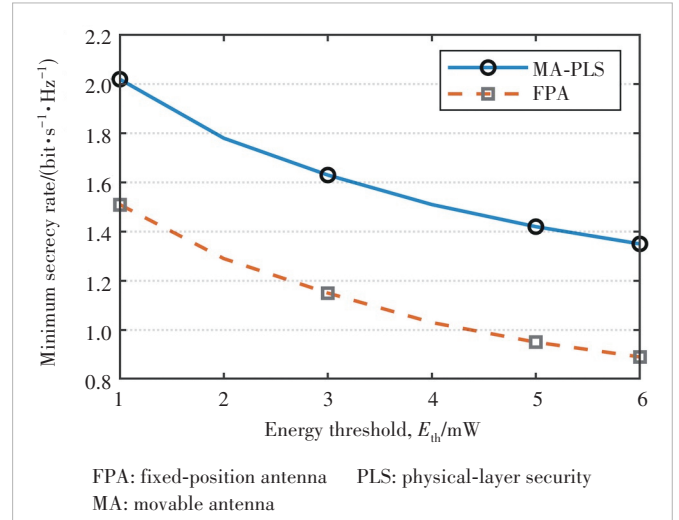


Figure 5. Minimum secrecy rate versus the minimum harvested-energy threshold of ERs

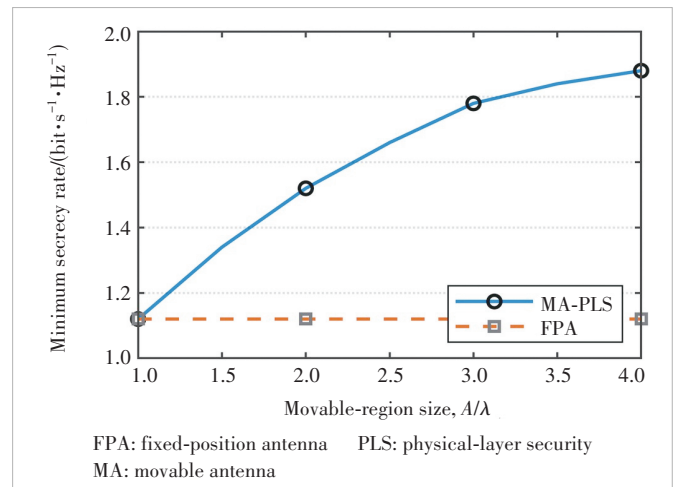


Figure 6. Minimum secrecy rate versus the normalized movable-region side length A/λ

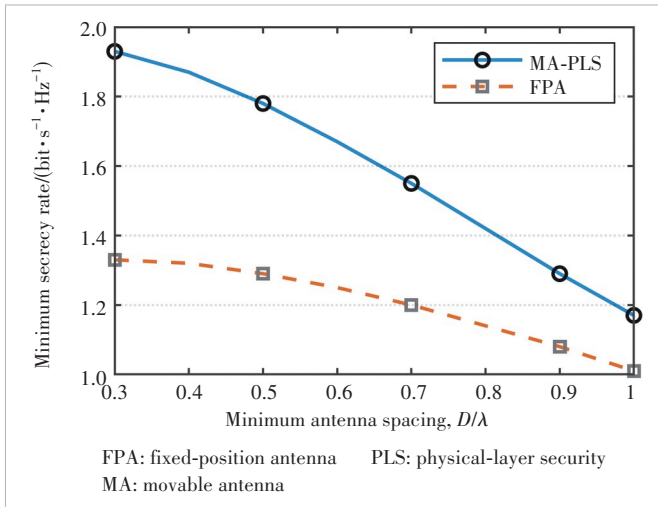


Figure 7. Minimum secrecy rate versus the normalized minimum antenna distance D/λ under different schemes

sults showed that MA position optimization and energy-signal reuse can effectively improve the minimum secrecy rate compared with fixed-position antenna benchmarks while satisfying ER energy requirements. Future work will incorporate MA movement-energy consumption and mechanical delay and develop low-complexity online resource-allocation algorithms for large-scale time-varying secure IDET networks.

References

- [1] Clerckx B, Zhang R, Schober R, et al. Fundamentals of wireless information and power transfer: from RF energy harvester models to signal and system designs [J]. *IEEE journal on selected areas in communications*, 2019, 37(1): 4 – 33. DOI: 10.1109/JSAC.2018.2872615
- [2] Bi S, Ho C K, Zhang R. Wireless powered communication: opportunities and challenges [J]. *IEEE communications magazine*, 2015, 53(4): 117 – 125. DOI: 10.1109/MCOM.2015.7081084
- [3] Hu J, Yang K, Wen G J, et al. Integrated data and energy communication network: a comprehensive survey [J]. *IEEE communications surveys & tutorials*, 2018, 20(4): 3169 – 3219. DOI: 10.1109/COMST.2018.2860778
- [4] Wyner A D. The wire-tap channel [J]. *Bell system technical journal*, 1975, 54(8): 1355 – 1387. DOI: 10.1002/j.1538-7305.1975.tb02040.x
- [5] Goel S, Negi R. Guaranteeing secrecy using artificial noise [J]. *IEEE transactions on wireless communications*, 2008, 7(6): 2180 – 2189. DOI: 10.1109/TWC.2008.060848
- [6] Ng D W K, Lo E S, Schober R. Robust beamforming for secure communication in systems with wireless information and power transfer [J]. *IEEE transactions on wireless communications*, 2014, 13(8): 4599 – 4615. DOI: 10.1109/TWC.2014.2314654
- [7] Zhu L P, Ma W Y, Zhang R. Movable antennas for wireless communication: opportunities and challenges [J]. *IEEE communications magazine*, 2024, 62(6): 114 – 120. DOI: 10.1109/MCOM.001.2300212
- [8] Zhu L P, Ma W Y, Zhang R. Modeling and performance analysis for movable antenna enabled wireless communications [J]. *IEEE transactions on wireless communications*, 2024, 23(6): 6234 – 6250. DOI: 10.1109/TWC.2023.3330887
- [9] Ma W Y, Zhu L P, Zhang R. MIMO capacity characterization for movable antenna systems [J]. *IEEE transactions on wireless communications*, 2024, 23(4): 3392 – 3407. DOI: 10.1109/TWC.2023.3307696
- [10] Zhao Y Z, Zhang L, Yang H, et al. Reconfigurable antennas for next-generation mobile communication networks: a comprehensive survey and tutorial [J]. *IEEE communications surveys & tutorials*, 2026, 28: 5267 – 5306. DOI: 10.1109/COMST.2026.3673688
- [11] Zhu L P, Ma W Y, Ning B Y, et al. Movable-antenna enhanced multiuser communication via antenna position optimization [J]. *IEEE transactions on wireless communications*, 2024, 23(7): 7214 – 7229. DOI: 10.1109/TWC.2023.3338626
- [12] Ma W Y, Zhu L P, Zhang R. Compressed sensing based channel estimation for movable antenna communications [J]. *IEEE communications letters*, 2023, 27(10): 2747 – 2751. DOI: 10.1109/LCOMM.2023.3310535
- [13] Xiao Z Y, Cao S Q, Zhu L P, et al. Channel estimation for movable antenna communication systems: a framework based on compressed sensing [J]. *IEEE transactions on wireless communications*, 2024, 23(9): 11814 – 11830. DOI: 10.1109/TWC.2024.3385110
- [14] Zhao Y Z, Zhang L, Hu J, et al. Reflective index modulation for IRS assisted integrated data and energy transfer [J]. *IEEE transactions on wireless communications*, 2024, 23(9): 11508 – 11520. DOI: 10.1109/TWC.2024.3382578
- [15] Huang Q X, Hu J, Zhao Y Z, et al. Holographic integrated data and energy transfer [J]. *IEEE transactions on wireless communications*, 2024, 23(12): 18987 – 19002. DOI: 10.1109/TWC.2024.3477750
- [16] Xiang P R, Zhao Y Z, Zhao Y H, et al. Joint resource allocation for movable antenna-assisted integrated data and energy transfer [J]. *IEEE wireless communications letters*, 2026, 15: 2704 – 2708. DOI: 10.1109/LWC.2026.3684328
- [17] Sun H, Zhu L, Mei W D, et al. Power measurement-based channel estimation for IRS-enhanced wireless coverage [J]. *IEEE transactions on wireless communications*, 2024, 23(12): 19183 – 19198. DOI: 10.1109/TWC.2024.3480314
- [18] Sun H, Zhu L, Xu J, et al. Channel gain map estimation based on 3-D virtual scatterer model [J]. *IEEE transactions on wireless communications*, 2026, 25: 15741 – 15757. DOI: 10.1109/TWC.2026.3685411
- [19] Sun H, Zhu L, Yan G, et al. Quasi-static IRS reflection design for network coverage enhancement: a practical RSRP-based approach [J]. *IEEE wireless communications*, 2026, early access. DOI: 10.1109/MWC.2026.3692436
- [20] Luo Z Q, Ma W K, So A M C, et al. Semidefinite relaxation of quadratic optimization problems [J]. *IEEE signal processing magazine*, 2010, 27(3): 20 – 34. DOI: 10.1109/MSP.2010.936019
- [21] Sidiropoulos N D, Davidson T N, Luo Z Q. Transmit beamforming for physical-layer multicasting [J]. *IEEE transactions on signal processing*, 2006, 54(6): 2239 – 2251. DOI: 10.1109/TSP.2006.872578

Biographies

Xiang Pengrong received her MS degree from the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC) in 2026. Her research interests include movable-antenna-assisted wireless communications and integrated data energy transfer.

Li Zhiguang is pursuing his MS degree at the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC). His research interests include movable-antenna-assisted wireless communications and integrated data energy transfer.

Zhao Yizhe (yzzhao@uestc.edu.cn) is currently an associate professor at University of Electronic Science and Technology of China (UESTC), where he received his PhD in information and communication engineering in 2021. He has been a visiting researcher with the Department of Electrical and Computer Engineering, University of California, Davis, USA. He is a member of IEEE and a senior member of China Institute of Communications. He is selected in Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST). He serves for *China Communications* and *Journal of Communications and Information Networks (JCIN)* as the Guest Editor, and is also a TPC member of several prestigious IEEE conferences, such as IEEE ICC and Globecom. He was the recipient of IEEE CSE Best Paper Award in 2023. His research interests include modulation and coding design, integrated data and energy transfer, and fluid antenna systems.



Robust Movable Antenna Position Optimization Under Imperfect Channel Path Parameters

Jia Qiaoran¹, Mei Weidong², Wei Xin², Chen Zhi²,
Ning Boyu²

(1. Glasgow College, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China, Chengdu 611731, China)

DOI: 10.12142/ZTECOM.202603006

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260826.1505.006.html>,
published online August 26, 2026

Manuscript received: 2026-06-10

Abstract: Movable antenna (MA) systems can reshape the wireless channel by repositioning antenna elements in a confined region, but most existing MA optimization methods depend on perfect channel state information (CSI). This paper studies robust joint transmit beamforming and antenna-position optimization for a downlink multiple-input single-output (MISO) MA system with imperfect CSI. Instead of treating CSI errors as unstructured channel perturbations, we introduce bounded uncertainties directly into physical multipath parameters, including path-gain amplitudes and angular parameters. The resulting worst-case signal-to-noise ratio (SNR) design is approximated through discrete sampling and a harmonic-mean surrogate that penalizes weak channel realizations. An alternating optimization algorithm is then developed, where transmit beamforming is updated on the complex sphere manifold, and antenna positions are optimized via a sequential grid search. Simulation results show that the proposed robust design yields higher SNR performance than non-robust designs and the fixed-position antenna (FPA) baseline.

Keywords: alternating optimization; harmonic mean; imperfect channel state information; movable antenna; robust optimization

Citation (Format 1): Jia Q R, Mei W D, Wei X, et al. Robust movable antenna position optimization under imperfect channel path parameters [J]. *ZTE Communications*, 2026, 24(3): 40 – 47. DOI: 10.12142/ZTECOM.202603006

Citation (Format 2): Q. R. Jia, W. D. Mei, X. Wei, et al., “Robust movable antenna position optimization under imperfect channel path parameters,” *ZTE Communications*, vol. 24, no. 3, pp. 40 – 47, Sept. 2026. doi: 10.12142/ZTECOM.202603006.

1 Introduction

The rapid growth of sixth generation (6G) and next-generation Wi-Fi services is pushing wireless systems toward higher reliability, higher capacity, and finer spatial control. Conventional multiple-input multiple-output (MIMO) systems improve spatial resolution by deploying more fixed antennas, but continuous scaling of static arrays increases radio-frequency hardware cost, energy consumption, and site constraints. Movable antenna (MA) technology has emerged as a complementary paradigm to conventional fixed-antenna architectures. Instead of solely increasing the number of fixed radiators, MA systems allow antenna elements to be repositioned within a finite local region. Such reconfigurability introduces additional spatial degrees of freedom, enabling the transmitter to better exploit fine-scale electromagnetic-field variations and mitigate unfavorable fading effects^[1–9].

Existing MA research has established field-response channel modeling, capacity characterization, and position optimi-

zation methods for both continuous and sampled movement regions^[8–12]. Recent studies have further expanded MA applications to a wide range of wireless scenarios, including multi-beam coverage, spectrum sharing, and integrated sensing and communication^[13–19]. In parallel, hardware-aware MA designs have been investigated by considering movement delay, mechanical energy consumption, and channel-map estimation^[20–22]. A recent survey has further provided a comprehensive overview of MA and other reconfigurable antenna architectures for next-generation mobile communication networks^[23].

However, a major practical challenge persists, as existing MA optimization frameworks mostly rely on the assumption of perfect channel state information (CSI). In realistic systems, CSI is inevitably degraded by estimation errors, feedback quantization, calibration mismatch, and tracking latency. Since MA performance is highly dependent on the phase relationship between multipath components and antenna coordinates, even small amplitude or angular errors may shift the ap-

parent spatial peaks of the channel field. Consequently, an antenna layout optimized for the estimated channel may suffer significant performance degradation when the actual physical parameters deviate from their nominal values. Robust MA position optimization under norm-bounded and Gaussian channel-vector errors has recently been studied in Ref. [24]. More broadly, robust designs for conventional wireless systems commonly adopt either deterministic norm-bounded uncertainty sets to guarantee performance for all admissible errors^[25], or statistical CSI-error models that control outage probability through chance constraints and tractable conservative approximations^[26]. A globally optimal discrete MA design considering imperfect CSI and power consumption was also investigated in Ref. [27]. These studies establish the importance of CSI-robust transmission and MA position optimization; however, their uncertainty models are mainly imposed on the overall channel responses at each antenna position and do not explicitly preserve the structured dependence of the MA channel on physical path gains and angles.

To address this issue, this paper focuses on robust antenna position optimization with CSI uncertainty of the physical multipath parameters. Specifically, bounded uncertainty sets are considered for path-gain amplitudes, elevation angles, and azimuth angles. Under this structured uncertainty model, we formulate a robust joint beamforming and antenna-position optimization problem for MA-enhanced downlink transmission. The design objective is to maximize the worst-case signal-to-noise ratio (SNR) over uncertain channel realizations. However, the resulting optimization problem is highly non-convex and difficult to solve. To tackle this challenge, we propose a harmonic-mean-based surrogate objective to approximate the finite-scenario worst-case SNR. Furthermore, we develop an alternating optimization algorithm that integrates Riemannian beamforming optimization and sequential grid-based search to optimize the transmit beamforming and antenna positions, respectively. Finally, simulation results demonstrate that the proposed robust design provides notable performance gains over the non-robust MA designs and the fixed-position antenna (FPA) benchmark.

2 System Model and Robust Formulation

2.1 Field-Response Channel Model

Consider a downlink communication system in which a base station equipped with N movable antennas serves a single user terminal equipped with an FPA, as shown in Fig. 1. The n -th MA is located at $\mathbf{t}_n = [x_n, y_n]^T$ in a two-dimensional movement region $\mathcal{C}_t$. The aggregate antenna-position matrix is $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_N]$. To maintain physical feasibility and reduce severe mutual coupling, each antenna must remain in $\mathcal{C}_t$ and any two antennas must satisfy $\|\mathbf{t}_n - \mathbf{t}_m\|_2 \geq D_{\min}$ for $n \neq m$, where $D_{\min}$ represents the minimum allowable antenna spacing. The

transmit beamforming vector is denoted by $\mathbf{w}$ in $\mathbb{C}^{N \times 1}$, with $\|\mathbf{w}\|_2^2 = 1$, and the transmit power is denoted by P_t .

Under the field-response model, local antenna movement changes the received phase of each multipath component while the large-scale path parameters remain approximately fixed within the small movement region. For L propagation paths, the channel vector from the base station (BS) to the user is modeled as:

$$\mathbf{h}(\mathbf{T}) = \sum_{l=1}^L g_l \mathbf{a}(\mathbf{T}, \theta_l, \phi_l) \quad (1),$$

where g_l is the complex gain of the l -th path, and θ_l and ϕ_l are its elevation and azimuth parameters. The array response $\mathbf{a}(\mathbf{T}, \theta_l, \phi_l)$ contains N unit-modulus components, and the phase of the n -th component is:

$$\psi_n(\theta_l, \phi_l) = \frac{2\pi}{\lambda} (x_n \cos \theta_l \sin \phi_l + y_n \sin \theta_l) \quad (2),$$

where λ denotes the wavelength.

The received signal is expressed as $y = \sqrt{P_t} \mathbf{h}^H(\mathbf{T}) \mathbf{w} s + n$, where s is a unit-power information symbol satisfying $\mathbb{E}[|s|^2] = 1$, and $n \sim \mathcal{CN}(0, \sigma^2)$ is additive white Gaussian noise, where σ^2 denotes the average noise power.

The received SNR is:

$$\gamma(\mathbf{w}, \mathbf{T}) = \frac{P_t}{\sigma^2} \mathbf{w}^H \mathbf{H}(\mathbf{T}) \mathbf{w} \quad (3),$$

where $\mathbf{H}(\mathbf{T}) = \mathbf{h}(\mathbf{T}) \mathbf{h}^H(\mathbf{T})$.

2.2 Structured CSI Uncertainty and Robust Objective

In conventional robust beamforming, the channel is often modeled as $\mathbf{h} = \hat{\mathbf{h}} + \Delta \mathbf{h}$ with $\|\Delta \mathbf{h}\|_2 \leq \epsilon$. Although convenient, this unstructured perturbation does not capture how angular or amplitude errors affect a position-dependent field response. We therefore model uncertainty directly through the

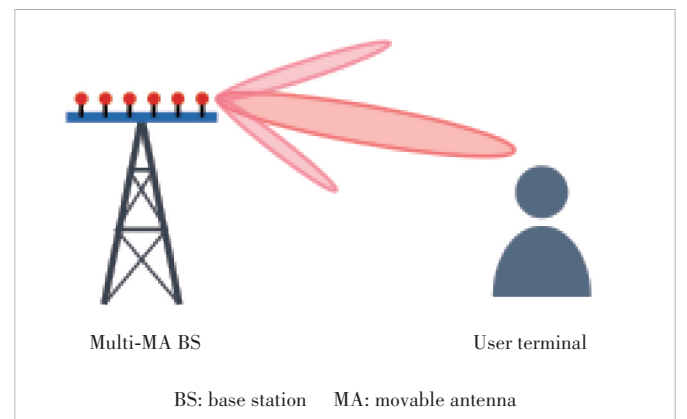


Figure 1. Downlink MA-enhanced communication system

physical path parameters. Let $g_l = \rho_l \exp(j\alpha_l)$, where ρ_l is the path-gain amplitude and α_l is the nominal gain phase. The amplitude and angular parameters obey bounded intervals, i.e.,

$$\rho_l \in [\rho_l^{\min}, \rho_l^{\max}], \theta_l \in [\theta_l^{\min}, \theta_l^{\max}], \phi_l \in [\phi_l^{\min}, \phi_l^{\max}], \forall l \quad (4).$$

We assume that the gain phase is kept at its nominal value. This assumption focuses the study on the dominant amplitude and angular uncertainty while preserving the deterministic field-response structure. For a given $\mathbf{T}$, we define an uncertainty set $\Lambda(\mathbf{T})$ that consists of all possible channel covariance matrices $\mathbf{H}(\mathbf{T})$ generated by parameters satisfying Eq. (4).

The robust design objective is to maximize the worst-case SNR over the structured uncertainty set:

$$(P1) \quad \begin{aligned} & \max_{\mathbf{w}, \mathbf{T}} \min_{\mathbf{H}(\mathbf{T}) \in \Lambda(\mathbf{T})} \frac{P_t}{\sigma^2} \mathbf{w}^H \mathbf{H}(\mathbf{T}) \mathbf{w} \\ & \text{s.t.} \quad \mathbf{t}_n \in \mathcal{C}_t, \forall n, \\ & \quad \|\mathbf{t}_n - \mathbf{t}_m\|_2 \geq D_{\min}, n \neq m, \\ & \quad \|\mathbf{w}\|_2^2 = 1. \end{aligned}$$

However, (P1) is difficult to handle because the uncertainty set is continuous; the channel depends nonlinearly on the antenna coordinates, and $\mathbf{w}$ and $\mathbf{T}$ are coupled. Therefore, a tractable approximation is needed.

3 Harmonic-Mean Robust Optimization

3.1 Discretization and Surrogate Objective

The continuous uncertainty set is first approximated by a finite set of representative discrete points. For each path, the amplitude, elevation angle, and azimuth angle are uniformly sampled inside their intervals. To avoid an exorbitant number of discrete points, a coupled sampling strategy is adopted, where a common index is used for amplitude, elevation, and azimuth perturbations, and each index triplet is applied jointly to all L paths. Specifically, let $\xi_s = \{g_l^{(s)}, \theta_l^{(s)}, \phi_l^{(s)}\}_{l=1}^L$ denote the S -th sampled parameter triple, and S denote the total number of sampled parameter triples. The corresponding channel vector and covariance matrix are:

$$\mathbf{h}_s(\mathbf{T}) = \sum_{l=1}^L g_l^{(s)} \mathbf{a}(\mathbf{T}, \theta_l^{(s)}, \phi_l^{(s)}), \quad \mathbf{H}_s(\mathbf{T}) = \mathbf{h}_s(\mathbf{T}) \mathbf{h}_s^H(\mathbf{T}) \quad (5).$$

For the s -th representative channel realization, we define the received channel gain as $x_s(\mathbf{w}, \mathbf{T}) = \mathbf{w}^H \mathbf{H}_s(\mathbf{T}) \mathbf{w}$.

Because P_t/σ^2 is a positive constant independent of $\mathbf{w}$ and $\mathbf{T}$, it is omitted in the surrogate objective without loss of optimality. A direct robust objective would maximize the minimum of $x_s(\mathbf{w}, \mathbf{T})$ over all representative channel realizations. However, the minimum operator is non-smooth. We therefore

adopt the following regularized harmonic mean:

$$\text{HM}_\delta = \left[\frac{1}{S} \sum_{s=1}^S \frac{1}{x_s(\mathbf{w}, \mathbf{T}) + \delta} \right]^{-1} \quad (6),$$

to approximate the minimum, where $\delta > 0$ is a small positive regularization constant introduced to prevent numerical singularities.

The harmonic mean is strongly affected by small elements, so it penalizes the parameter triple with poor channel quality more effectively than an arithmetic-mean objective. It is not identical to the strict minimum; instead, it provides a smooth robustness-promoting surrogate.

Given Eq. (6), the robust joint design problem is reformulated as (P2). Fig. 2 summarizes the resulting robust optimization flow.

$$(P2) \quad \begin{aligned} & \max_{\mathbf{w}, \mathbf{T}} \quad - \sum_{s=1}^S \frac{1}{\mathbf{w}^H \mathbf{H}_s(\mathbf{T}) \mathbf{w} + \delta} \\ & \text{s.t.} \quad \mathbf{t}_n \in \mathcal{C}_t, \forall n, \\ & \quad \|\mathbf{t}_n - \mathbf{t}_m\|_2 \geq D_{\min}, n \neq m, \\ & \quad \|\mathbf{w}\|_2^2 = 1. \end{aligned}$$

3.2 Alternating Optimization

The surrogate problem in (P2) remains non-convex. We therefore adopt an alternating optimization framework, in

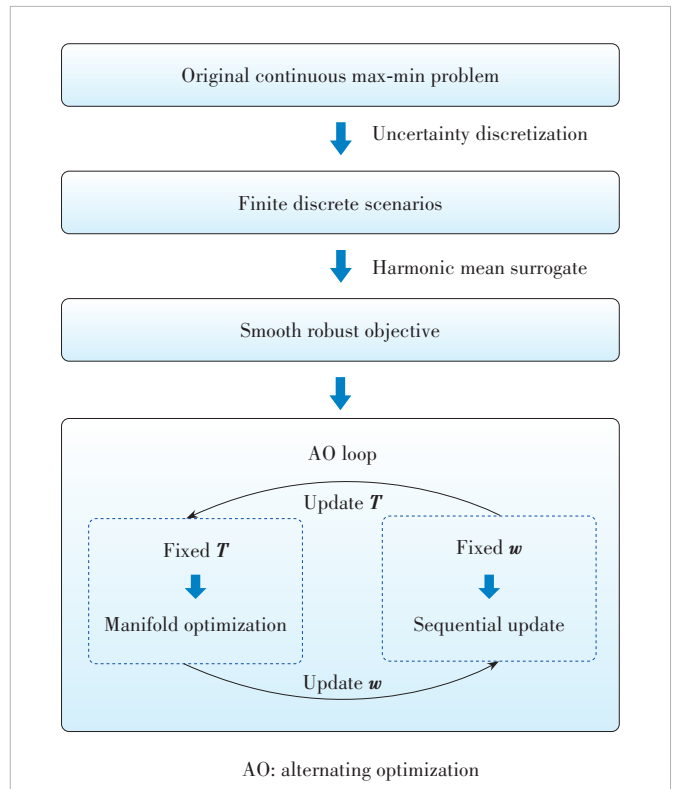


Figure 2. Robust joint beamforming and MA-position optimization flow

which the transmit beamforming vector and the antenna positions are optimized alternately while the other variable is kept fixed.

With fixed antenna positions $\mathbf{T}$, the beamforming vector is optimized on the complex unit sphere $\mathcal{M} = \{\mathbf{w} \in \mathbb{C}^N : \|\mathbf{w}\|_2 = 1\}$.

Recall that $x_s(\mathbf{w}|\mathbf{T}) = \mathbf{w}^H \mathbf{H}_s(\mathbf{T}) \mathbf{w}$ denotes the received channel power gain under the s -th representative channel realization. Since maximizing the harmonic-mean objective in Eq. (6) is equivalent to minimizing its reciprocal-sum denominator, the beamforming subproblem is implemented as:

$$\min_{\mathbf{w} \in \mathcal{M}} F(\mathbf{w}; \mathbf{T}) = \kappa \sum_{s=1}^S \frac{1}{x_s(\mathbf{w}, \mathbf{T}) + \delta} \quad (7),$$

where κ is a positive numerical scaling factor that does not affect the optimal beamforming vector, and it is set as 10^6 in the simulation.

The Euclidean gradient of $F(\mathbf{w}; \mathbf{T})$ supplied to the manifold solver is given by:

$$\nabla_{\mathbf{w}} F(\mathbf{w}; \mathbf{T}) = -2\kappa \sum_{s=1}^S \frac{\mathbf{H}_s(\mathbf{T}) \mathbf{w}}{[x_s(\mathbf{w}, \mathbf{T}) + \delta]^2} \quad (8).$$

The Euclidean gradient in Eq. (8) is projected onto the tangent space of the complex unit sphere to obtain the Riemannian gradient as $\text{grad } F(\mathbf{w}; \mathbf{T}) = \nabla_{\mathbf{w}} F(\mathbf{w}; \mathbf{T}) - \mathbf{w} \text{Re}\{\mathbf{w}^H \nabla_{\mathbf{w}} F(\mathbf{w}; \mathbf{T})\}$.

The beamforming subproblem is then solved using the Riemannian conjugate-gradient method implemented in Manopt, which performs the search-direction update, step-size selection, and retraction onto the complex unit sphere. The beamforming solution obtained in the preceding alternating optimization (AO) iteration is used as the initial point of the current beamforming update. The inner solver terminates when the cost reduction over two consecutive iterations is less than a prescribed value (set as 10^{-6} in the simulation).

With fixed beamforming $\mathbf{w}$, the antenna-position optimization subproblem is solved by sequential grid search. The movement region $\mathcal{C}_i$ is discretized into K candidate points $\mathbf{q}_k \in \mathcal{C}_i$, $k = 1, \dots, K$, and the resulting grid set is denoted by $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_K\}$.

During one antenna-position sweep, the antennas are updated sequentially from $n = 1$ to N . When optimizing the position of the n -th antenna, the positions of the other $N - 1$ antennas are temporarily fixed. The feasible candidate set is constructed as:

$$\mathcal{Q}_n^{\text{feas}}(\mathbf{T}_{-n}) = \left\{ \mathbf{q} \in \mathcal{Q} : \|\mathbf{q} - \mathbf{t}_m\|_2 \geq D_{\min}, \forall m \neq n \right\} \quad (9),$$

where $\mathbf{T}_{-n}$ denotes the positions of all antennas except the n -th antenna.

For each $\mathbf{q} \in \mathcal{Q}_n^{\text{feas}}(\mathbf{T}_{-n})$, let $\mathbf{T}_n(\mathbf{q})$ denote the antenna-

position matrix obtained by replacing $\mathbf{t}_n$ with $\mathbf{q}$, while keeping the other antenna positions unchanged. The channel matrices $\mathbf{H}_s(\mathbf{T}_n(\mathbf{q}))$, $s = 1, \dots, S$, are recalculated, and the reciprocal-sum objective $F(\mathbf{w}|\mathbf{T}_n(\mathbf{q}))$ is evaluated. The position of the n -th antenna is then updated according to:

$$\mathbf{t}_n \leftarrow \arg \min_{\mathbf{q} \in \mathcal{Q}_n^{\text{feas}}(\mathbf{T}_{-n}) \cup \{\mathbf{t}_n\}} F(\mathbf{w}; \mathbf{T}_n(\mathbf{q})) \quad (10).$$

Before evaluating the feasible grid candidates, the current position of the n -th antenna is retained as the incumbent solution. Therefore, if none of the feasible candidates improves the objective value, the antenna remains at its current position. Each updated antenna position is immediately used when constructing the feasible candidate set for the subsequent antenna. The outer loop of the alternating optimization alternates between the manifold beamforming update and the sequential antenna-position sweep until the relative objective change is below a tolerance or the maximum number of iterations is reached.

The complexity of the proposed algorithm is analyzed next. If I_w inner iterations are used for beamforming optimization, the beamforming update costs $\mathcal{O}(I_w SN^2)$. For the antenna position optimization, a full antenna-position sweep costs $\mathcal{O}(NKS)$. With I_{A0} outer iterations, the overall complexity is thus given by $\mathcal{O}(I_{A0} SN(I_w N + K))$.

4 Simulation Results

4.1 Setup and Benchmarks

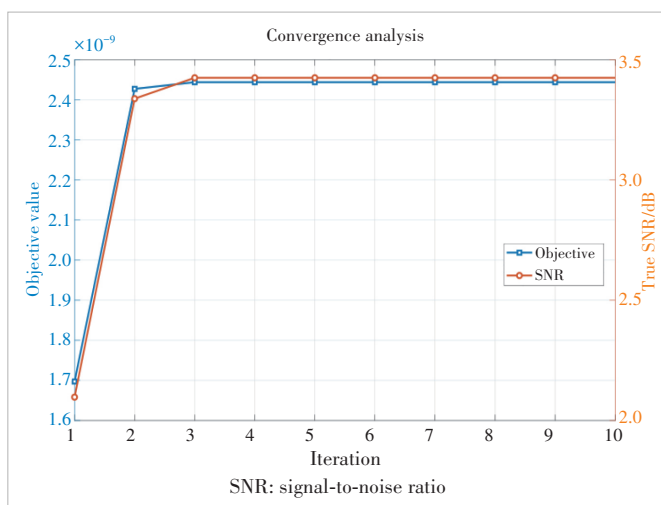
Numerical results are provided in this section to evaluate convergence, robustness against CSI uncertainty, and scalability with the number of antennas. Unless otherwise stated, the carrier frequency is 5 GHz, the wavelength is 0.06 m, the number of paths is $L = 3$, and the base-station-to-user distance is 50 m. The transmit region is a square of $4\lambda \times 4\lambda$, and the minimum antenna spacing is $\lambda/2$. Furthermore, the interval bounds in Eq. (4) are generated by introducing a scaling factor Δ . Specifically, let $\bar{\rho}_l$, $\bar{\theta}_l$, and $\bar{\phi}_l$ denote the estimated amplitude, elevation angle, and azimuth angle of the l -th path, respectively. The bounds in Eq. (4) are set as $\theta_l^{\min} = \bar{\theta}_l - \frac{\pi}{12} \Delta$, $\theta_l^{\max} = \bar{\theta}_l + \frac{\pi}{12} \Delta$, $\phi_l^{\min} = \bar{\phi}_l - \frac{\pi}{12} \Delta$, $\phi_l^{\max} = \bar{\phi}_l + \frac{\pi}{12} \Delta$, $\rho_l^{\min} = (1 - 0.1\Delta) \bar{\rho}_l$, and $\rho_l^{\max} = (1 + 0.1\Delta) \bar{\rho}_l$. It follows that a larger value of Δ produces wider amplitude and angular uncertainty intervals and represents more severe CSI mismatch. The main simulation parameters are listed in Table 1.

4.2 Convergence Behavior

First, Fig. 3 presents the convergence plot of the proposed alternating optimization algorithm in terms of the objective function and actual received SNR with $\Delta = 1$.

Table 1. Summary of simulation parameters

Parameter	Default Value
Carrier frequency (f_c)	5 GHz
Wavelength (λ)	0.06 m
Default number of MAs (N)	4
Number of paths (L)	3
BS-user distance (d)	50 m
Path-loss exponent (α)	2.8
Transmit region ($A \times A$)	$4\lambda \times 4\lambda$
Minimum spacing ($D_{\min}$)	$\lambda/2$
Transmit power (P_t)	46 dBm
Noise power (σ^2)	-44 dBm


Figure 3. Convergence behavior of the proposed alternating optimization algorithm

It is shown that the harmonic-mean surrogate objective increases rapidly at the early iterations and gradually reaches a stable plateau, verifying the effectiveness of the alternating updates. The actual received SNR follows a similar trend, which demonstrates that the surrogate objective provides a reliable optimization direction for performance improvement. Moreover, the smooth convergence behavior indicates favorable numerical stability, which is desirable for MA systems to avoid unnecessary antenna movement caused by oscilla-

tory updates.

4.3 Benchmark Schemes

To comprehensively evaluate the performance of the proposed algorithm, it is compared against three benchmark schemes in the following. These benchmark schemes are selected to distinguish the effects of three key factors: perfect versus imperfect CSI, movable versus fixed antenna architectures, and robust versus non-robust position optimization.

First, the Perfect-CSI+MA scheme assumes that true channel parameters are available during optimization and employs the graph-based dynamic programming algorithm (Graph-DP) proposed in Ref. [11] to optimize MA positions. However, since the conventional Graph-DP algorithm was originally developed for one-dimensional linear antenna arrangements, it cannot be directly applied to the two-dimensional planar deployment considered in this study due to the increased geometric complexity. To address this, we incorporate a greedy collision-resolution mechanism into the discrete dynamic programming procedure, ensuring that the minimum-distance constraint is satisfied while pursuing high field-response values. Its performance serves as an upper bound for the proposed scheme and other benchmark schemes.

Second, the Imperfect-CSI+MA (Graph-DP) scheme represents a conventional non-robust MA design. It employs the same two-dimensional Graph-DP implementation described above. However, instead of using the true channel parameters, this scheme is directly fed with the imperfectly estimated CSI.

Third, the Imperfect-CSI+FPA scheme acts as the traditional static-array baseline. It utilizes a fixed-position antenna array without spatial mobility. The estimated channel vector is constructed based on the imperfect path parameters, and standard maximum ratio transmission (MRT) precoding is then applied to align the beamforming vector with the estimated channel. For clarity, the main characteristics of the proposed method and all benchmark schemes are summarized in Table 2.

4.4 Robustness Against CSI Uncertainty

Fig. 4 plots the average actual received SNR versus the CSI error-bound scaling factor Δ , averaged over 100 independent Monte Carlo realizations. In each realization, the actual elevation and azimuth angles are independently and uniformly gen-

Table 2. Summary of the proposed method and benchmark schemes

Scheme	CSI Assumption	Antenna Type	Design Principle
Proposed robust MA	Imperfect CSI	Movable	Harmonic-mean robust AO
Perfect-CSI+MA	Perfect CSI	Movable	Graph-DP with true CSI
Imperfect-CSI+MA	Imperfect CSI	Movable	Graph-DP with estimated CSI
Imperfect-CSI+FPA	Imperfect CSI	Fixed	MRT with estimated CSI

AO: alternating optimization

FPA: fixed-position antenna

MA: movable antenna

CSI: channel state information

Graph-DP: graph-based dynamic programming algorithm

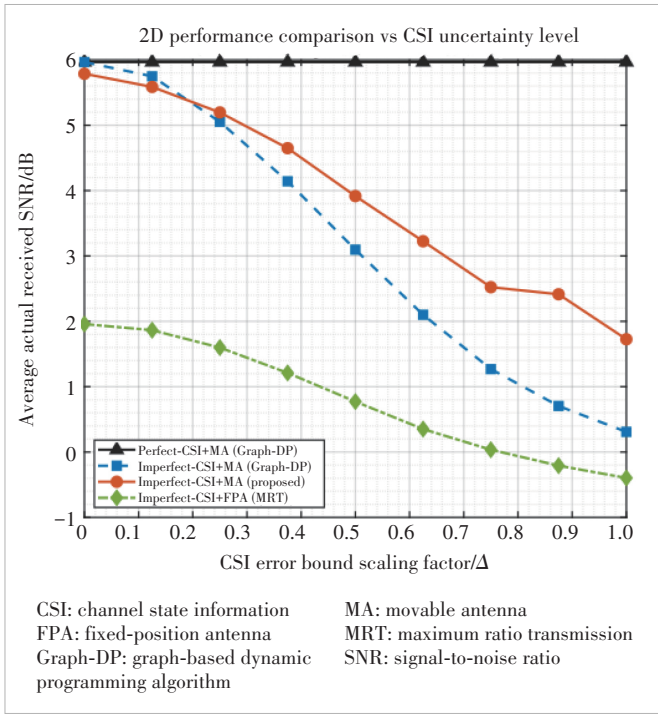


Figure 4. Average actual received SNR versus CSI uncertainty level

erated within $[0, \pi/2]$ and $[-\pi, \pi]$, respectively. The complex path gains are generated according to independent circularly symmetric complex Gaussian (CSCG) distributions and normalized based on the large-scale path loss. The estimated channel parameters are obtained by introducing bounded perturbations into the actual parameters. The elevation- and azimuth-angle errors are bounded by $(\pi/12)\Delta$, while the relative path-gain amplitude error is bounded by 0.1Δ . The path phases remain unchanged. Within each Monte Carlo realization, the perturbation directions are fixed as Δ varies, and only their magnitudes are scaled.

It should be noted that the worst-case SNR over the continuous uncertainty set in (P1) is also a widely adopted performance metric. However, in the proposed algorithm, the continuous uncertainty set is approximated by a finite number of representative channel realizations, and the non-smooth minimum operator is further replaced by the harmonic-mean surrogate in Eq. (6). Therefore, the proposed method promotes worst-case robustness by assigning greater penalties to weak channel realizations, rather than directly obtaining the exact worst-case SNR. In view of this, the average actual received SNR is adopted to evaluate the practical performance of the proposed scheme.

Several observations can be made from Fig. 4. First, the Perfect-CSI+MA (Graph-DP) scheme achieves the best SNR performance among the considered schemes since it optimizes the antenna positions using the true channel parameters. Second, the Imperfect-CSI+FPA (MRT) baseline exhibits the low-

est performance across the considered uncertainty range. The gap between the MA-based schemes and the FPA baseline confirms the fundamental benefit of antenna mobility: spatial movement provides additional degrees of freedom to avoid unfavorable fading, even when CSI is imperfect.

A clear performance crossover is observed between the proposed robust MA scheme and the Imperfect-CSI+MA (Graph-DP) benchmark. In the extremely low-uncertainty regime, the proposed robust design exhibits a marginal performance loss compared with the non-robust Graph-DP benchmark. This is expected because the Graph-DP scheme directly optimizes the antenna positions according to the estimated field-response peak when the nominal CSI is almost accurate. However, as the uncertainty level Δ increases, the non-robust Graph-DP scheme suffers from more rapid performance degradation, whereas the proposed robust design maintains a slower decay rate and achieves a clear SNR advantage under moderate and severe uncertainty.

4.5 Performance Scalability with Antenna Array Size

Next, we evaluate how the performance of the proposed scheme scales with the number of transmit antennas N . To isolate the impact of array size under imperfect CSI, this evaluation is conducted at a fixed moderate uncertainty level of $\Delta = 0.5$. In this setting, each uncertain parameter is discretized into 8 points in its interval, and the results are averaged over 100 independent Monte Carlo realizations.

Fig. 5 plots the average actual received SNR as a function of N . It is observed that all considered schemes achieve

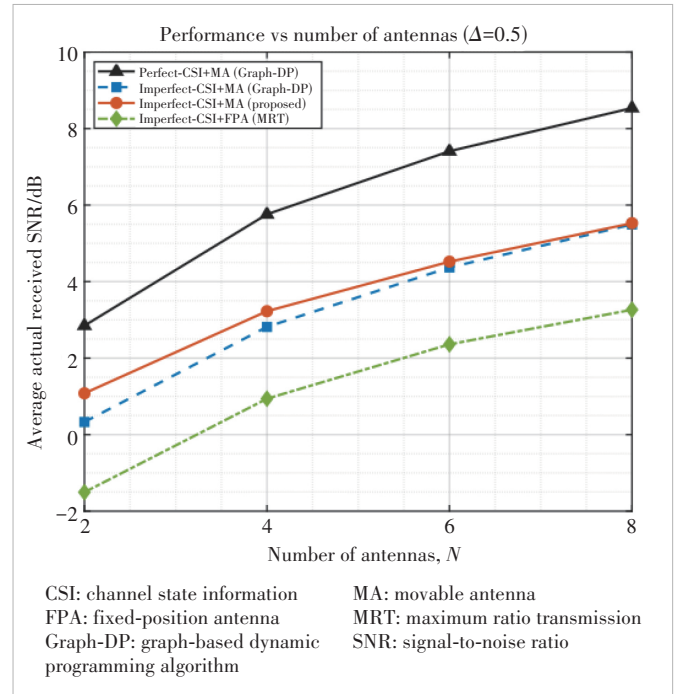


Figure 5. Average actual received SNR versus the number of transmit antennas at $\Delta=0.5$

higher received SNR as the number of antennas increases. This trend is consistent with the fact that additional antennas enhance the array gain and improve the signal-combining capability. As expected, the Perfect-CSI+MA (Graph-DP) scheme provides the best SNR performance among the considered schemes. In contrast, the Imperfect-CSI+FPA (MRT) baseline remains at the bottom due to the absence of spatial mobility. The substantial gap between the FPA baseline and the MA-based schemes further confirms that movable antennas provide an important spatial degree of freedom for combating fading and CSI mismatch. Furthermore, for smaller arrays, such as $N = 2$ and $N = 4$, the proposed robust design noticeably outperforms the non-robust Graph-DP benchmark. However, as N further increases, the performance gap between the two imperfect-CSI MA schemes gradually narrows. This is because the feasible movement space becomes increasingly crowded as N increases. The minimum-distance constraint restricts the permissible displacement of each antenna, forcing both robust and non-robust algorithms into similar dense geometric configurations. In this regime, the gain from position optimization becomes less dominant, while the array gain associated with more antennas becomes the primary source of SNR improvement.

5 Conclusions

This paper presented a robust joint beamforming and antenna-position optimization framework for MA-enhanced MISO downlink transmission under imperfect estimates of path parameters. The CSI uncertainty was embedded into physical multipath parameters, including path-gain amplitudes and angular parameters. To make the continuous worst-case SNR maximization problem tractable, the uncertainty region was sampled into finite parameter triples and optimized through a harmonic-mean-based surrogate. An alternating optimization algorithm was developed by combining Riemannian beamforming on the complex sphere and a sequential grid search for antenna positions.

Simulation results demonstrated stable convergence of the proposed robust MA optimization algorithm and showed that the proposed algorithm outperformed the non-robust MA benchmark when uncertainty increased. Future work will consider more practical aspects, including joint amplitude-phase-angle uncertainty, movement delay, mechanical energy consumption, mutual coupling, multi-user interference, and wide-band channel models.

References

- [1] Wang Z, Zhang J Y, Du H Y, et al. A tutorial on extremely large-scale MIMO for 6G: fundamentals, signal processing, and applications [J]. IEEE communications surveys & tutorials, 2024, 26(3): 1560 – 1605. DOI: 10.1109/comst.2023.3349276
- [2] Ning B Y, Tian Z B, Mei W D, et al. Beamforming technologies for ultra-massive MIMO in terahertz communications [J]. IEEE open journal of the communications society, 2023, 4: 614 – 658. DOI: 10.1109/ojcoms.2023.3245669
- [3] Chowdhury M Z, Shahjalal M, Ahmed S, et al. 6G wireless communication systems: applications, requirements, technologies, challenges, and research directions [J]. IEEE open journal of the communications society, 2020, 1: 957 – 975. DOI: 10.1109/OJCOMS.2020.3010270
- [4] Zhu L P, Zhang J, Xiao Z Y, et al. Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming [J]. IEEE transactions on wireless communications, 2019, 18(11): 5065 – 5079. DOI: 10.1109/twc.2019.2932070
- [5] Zhu L P, Ma W Y, Zhang R. Movable antennas for wireless communication: opportunities and challenges [J]. IEEE communications magazine, 2024, 62(6): 114 – 120. DOI: 10.1109/mcom.001.2300212
- [6] Ning B Y, Yang S J, Wu Y F, et al. Movable antenna-enhanced wireless communications: general architectures and implementation methods [J]. IEEE wireless communications, 2025, 32(5): 108 – 116. DOI: 10.1109/MWC.013.2400238
- [7] Wang M J, Gao F F, Jin S, et al. An overview of enhanced massive MIMO with array signal processing techniques [J]. IEEE journal of selected topics in signal processing, 2019, 13(5): 886 – 901. DOI: 10.1109/jstsp.2019.2934931
- [8] Ma W Y, Zhu L P, Zhang R. MIMO capacity characterization for movable antenna systems [J]. IEEE transactions on wireless communications, 2024, 23(4): 3392 – 3407. DOI: 10.1109/twc.2023.3307696
- [9] Zhu L P, Ma W Y, Zhang R. Modeling and performance analysis for movable antenna enabled wireless communications [J]. IEEE transactions on wireless communications, 2024, 23(6): 6234 – 6250. DOI: 10.1109/twc.2023.3330887
- [10] Zhu L P, Ma W Y, Ning B Y, et al. Movable-antenna enhanced multiuser communication via antenna position optimization [J]. IEEE transactions on wireless communications, 2024, 23(7): 7214 – 7229. DOI: 10.1109/twc.2023.3338626
- [11] Mei W D, Wei X, Ning B Y, et al. Movable-antenna position optimization: a graph-based approach [J]. IEEE wireless communications letters, 2024, 13(7): 1853 – 1857. DOI: 10.1109/lwc.2024.3393135
- [12] Liu C H, Mei W D, Chen Z, et al. A general optimization framework for movable antenna systems via discrete sampling [J]. IEEE wireless communications letters, 2026, 15: 475 – 479. DOI: 10.1109/lwc.2025.3629978
- [13] Ma W Y, Zhu L P, Zhang R. Multi-beam forming with movable-antenna array [J]. IEEE communications letters, 2024, 28(3): 697 – 701. DOI: 10.1109/lcomm.2024.3350647
- [14] Wang D, Mei W D, Ning B Y, et al. Movable antenna enhanced wide-beam coverage: joint antenna position and beamforming optimization [J]. IEEE transactions on wireless communications, 2026, 25: 3541 – 3558. DOI: 10.1109/twc.2025.3604598
- [15] Wei X, Mei W D, Wang D, et al. Joint beamforming and antenna position optimization for movable antenna-assisted spectrum sharing [J]. IEEE wireless communications letters, 2024, 13(9): 2502 – 2506. DOI: 10.1109/lwc.2024.3421636
- [16] An B N, Mei W D, Liu Y W, et al. STAR-RIS-enabled multi-path beam routing with passive beam splitting [J]. IEEE transactions on vehicular technology, 2026, 75(6): 11035 – 11050. DOI: 10.1109/tvt.2025.3650199
- [17] Wei X, Mei W D, Wu Q Q, et al. Movable antennas meet intelligent reflecting surface: friends or foes? [J]. IEEE transactions on communications, 2025, 73(11): 12756 – 12770. DOI: 10.1109/tcomm.2025.3588579
- [18] Wang Y S, Mei W D, Wei X, et al. Antenna position optimization for movable antenna-empowered near-field sensing [C]//International Conference on Communications Workshops (ICC Workshops). IEEE, 2025: 324 – 329. DOI: 10.1109/ICCWorkshops67674.2025.11162470
- [19] Lyu W T, Yang S J, Xiu Y, et al. Movable antenna enabled integrated sensing and communication [J]. IEEE transactions on wireless communications, 2025, 24(4): 2862 – 2875. DOI: 10.1109/twc.2025.3525631

- [20] Li Q L, Mei W D, Zhang R, et al. Trajectory optimization for minimizing movement delay in movable antenna systems [J]. *IEEE transactions on wireless communications*, 2026, 25: 6986 – 6999. DOI: 10.1109/twc.2025.3627982
- [21] Wei X, Mei W D, Huang X, et al. Mechanical power modeling and energy efficiency maximization for movable antenna systems [C]//*IEEE Global Communications Conference*. IEEE, 2025: 6117 – 6122. DOI: 10.1109/GLOBECOM59602.2025.11432508
- [22] Huang Y T, Mei W D, Wei X, et al. CNN-based channel map estimation for movable antenna systems [C]//*Proceedings of IEEE 26th International Workshop on Signal Processing and Artificial Intelligence for Wireless Communications (SPAWC)*. IEEE, 2025: 1 – 5. DOI: 10.1109/spawc66079.2025.11143328
- [23] Zhao Y Z, Zhang L, Yang H, et al. Reconfigurable antennas for next-generation mobile communication networks: a comprehensive survey and tutorial [J]. *IEEE communications surveys & tutorials*, 2026, 28: 5267 – 5306. DOI: 10.1109/COMST.2026.3673688
- [24] Ma H F, Mei W D, Wei X, et al. Robust movable-antenna position optimization with imperfect CSI for MISO systems [J]. *IEEE communications letters*, 2025, 29(7): 1594 – 1598. DOI: 10.1109/lcomm.2025.3569177
- [25] Vucic N, Boche H. Robust QoS-constrained optimization of downlink multiuser MISO systems [J]. *IEEE transactions on signal processing*, 2009, 57(2): 714 – 725. DOI: 10.1109/tsp.2008.2008553
- [26] Wang K Y, So A M, Chang T H, et al. Outage constrained robust transmit optimization for multiuser MISO downlinks: tractable approximations by conic optimization [J]. *IEEE transactions on signal processing*, 2014, 62(21): 5690 – 5705. DOI: 10.1109/TSP.2014.2354312
- [27] Wu Y F, Xu D F, Wing Kwan Ng D, et al. Globally optimal movable antenna-enabled multiuser communication: discrete antenna positioning, power consumption, and imperfect CSI [J]. *IEEE transactions on communications*, 2025, 73(10): 9903 – 9923. DOI: 10.1109/tcomm.2025.3582717

Biographies

Jia Qiaoran received his BS degree in communication engineering from Glasgow College, University of Electronic Science and Technology of China. He is currently pursuing his MS degree with the School of Electronics, Peking University, China. His research interests include movable antennas and wireless communications. He was the co-recipient of the IEEE ComSoc SPCC Technical

Committee Student Challenge and Video Contest Award in 2024.

Mei Weidong (wmei@uestc.edu.cn) received his BE degree in communication engineering and ME degree in communication and information systems from the University of Electronic Science and Technology of China (UESTC), in 2014 and 2017, respectively, and his PhD degree from the NUS Graduate School, National University of Singapore in 2021. He is currently a professor with UESTC. His research interests include reconfigurable MIMO, intelligent reflecting surfaces, wireless drone communications, and convex optimization techniques. Prof. Mei was the recipient of the Best Paper Award from the IEEE International Conference on Communications in 2021, and the Outstanding Master's Thesis Award from the Chinese Institute of Electronics in 2017. He serves as an associate editor for the *IEEE Transactions on Communications*, *IEEE Internet of Things Journal*, *IEEE Wireless Communications Letters* and *IEEE Open Journal of the Communications Society*.

Wei Xin received his BS degree in communication engineering from Chongqing University of Posts and Telecommunications, China in 2023. He is currently pursuing his MS degree in information and communication engineering with the National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China. His research interests include movable antennas, wireless drone communications, and convex optimization. He was the co-recipient of the IEEE ComSoc SPCC Technical Committee Student Challenge and Video Contest Award in 2024.

Chen Zhi received his BE, ME, and PhD degrees in electrical engineering from the University of Electronic Science and Technology of China (UESTC) in 1997, 2000, and 2006, respectively. In April 2006, he joined the National Key Laboratory of Science and Technology on Communications, UESTC, where he has been a professor since 2013. He was a visiting scholar with the University of California, Riverside, USA from 2010 to 2011. He is also the Deputy Director of the Key Laboratory of Terahertz Technology, Ministry of Education of the People's Republic of China. His current research interests include terahertz communication, 5G mobile communications, and tactile Internet. He serves as an editor of *IEEE Transactions on Communications*.

Ning Boyu received his BS degree in communication engineering from the Yingcai Honors College, University of Electronic Science and Technology of China (UESTC) in 2018, and his PhD degree from the National Key Laboratory of Wireless Communications, UESTC in 2023. From 2022 to 2023, he was a visiting student with the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include terahertz communication, movable antennas, intelligent reflecting surfaces, massive MIMO, physical-layer security, and convex optimization.



High Resolution Automotive Radar Imaging Integrating Doppler Division Multiple Access and Frequency Agility

Zhang Wei¹, Zhang Chenyong², Wang Yu², Ding Jin¹,
Li Yao²

(1. School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. Chengdu Nanoradar Technology Co., Ltd., Chengdu 610041, China)

DOI: 10.12142/ZTECOM.202603007

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260824.1839.002.html>,
published online August 25, 2026

Manuscript received: 2024-10-24

Abstract: Millimeter-wave (MMW) radar is one of the indispensable sensors for advanced driver assistance systems (ADASs) and autonomous driving. This paper adopts inter-pulse frequency agility (FA) to synthesize a broadband signal, achieving broadband imaging without increasing the sampling rate. Meanwhile, given that Doppler division multiple access (DDMA) has higher transmission efficiency than time division multiple access (TDMA) and is widely used in automotive radar, this paper proposes a signal transmission framework that integrates DDMA and FA. Simulation results demonstrate that the proposed DDMA FA multiple-input multiple-output (DDMA-FA-MIMO) processing architecture can yield 30 to 40 times more detection points for a single target compared to traditional methods, preliminarily achieving target contour imaging.

Keywords: automotive radar; radar imaging; stepped frequency radar; 4D radar; DDMA-FA-MIMO

Citation (Format 1): Zhang W, Zhang C Y, Wang Y, et al. High resolution automotive radar imaging integrating Doppler division multiple access and frequency agility [J]. *ZTE Communications*, 2026, 24(3): 48 – 58. DOI: 10.12142/ZTECOM.202603007

Citation (Format 2): W. Zhang, C.Y. Zhang, Y. Wang, et al., “High resolution automotive radar imaging integrating Doppler division multiple access and frequency agility,” *ZTE Communications*, vol. 24, no. 3, pp. 48 – 58, Sept. 2026. doi: 10.12142/ZTECOM.202603007.

1 Introduction

Advanced driver assistance systems (ADAS) have emerged as standard equipment in most automobiles, while autonomous driving technology is advancing rapidly^[1]. These systems rely on diverse sensors, including radar, light detection and ranging (LiDAR), and cameras, to assist the driver during driving and parking. Notably, millimeter-wave (MMW) radar exhibits significant advantages in environmental adaptability compared to other sensors like LiDAR and cameras, thus playing a pivotal role in automotive autonomous and assisted driving applications^[2]. Furthermore, in addition to measuring crucial parameters like the distance, speed, azimuth and elevation of targets, current 4D radars also can provide the capability of high-resolution point cloud imaging^[3].

However, the density of point clouds in 4D radar images is often sparse, and the superior angular resolution of LiDAR poses significant challenges to MMW radar. Given the limited aperture size, it is commonly perceived that inferior angular

resolution may hinder MMW radar's potential to play a significant role in future autonomous driving applications. Notably, Tesla's assisted driving solutions once excluded MMW radar from their systems, though it has been reintroduced in its latest iterations^[4].

The indispensability of MMW radar in automotive driving solutions hinges significantly on its ability to notably enhance resolution. This has emerged as a prominent research focus in the field. To enhance the resolution of MMW radar, extensive research focused on high-resolution and high-precision direction-of-arrival (DOA) has been conducted. Early approaches to DOA estimation relied on digital beamforming (DBF), whose angular resolution is limited by the physical array aperture, i.e., the Rayleigh limit^[5]. Increasing the antenna aperture is undoubtedly one of the most direct ways. However, considering the costs and installation difficulties, increasing the physical array aperture may not always be the most feasible solution. Fortunately, multiple-input multiple-output (MIMO) technology^[6–8] offers the potential to create a large virtual aperture without increasing costs dramatically. Specifically, time-division multiplexing (TDM) MIMO activates only one transmit antenna element at a time and establishes signal orthogonality in the time domain. Consequently, most commer-

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62371093 and 62231006.

cial automotive MIMO radar devices on the market, such as Texas Instruments' (TI) AWR1843 and AWR2243^[9], and NXP Semiconductors' TEF82xx and SAF85XX^[10], are based on TDM MIMO. These devices further support multi-chip cascading to enhance the aperture. For instance, the Continental ARS540 achieves an angular resolution of 1.2° through cascading four chips^[11]. However, it is noteworthy that a resolution of around 1° is still insufficient, resulting in sparse point clouds that are unable to form a distinct target contour to effectively support autonomous driving. The sparsity of these point clouds poses challenges in accurately perceiving and identifying objects in the environment, which is crucial for safe and reliable autonomous driving. Similarly, synthetic aperture radar (SAR)^[12-14] technologies can also increase the aperture by integrating received signals from diverse positions. However, motion errors introduce significant phase errors into received signals, leading to severe degradation in system performance. Additionally, the efficacy of SAR in forward-looking applications is considerably constrained.

Resolution can also be improved by employing super-resolution algorithms, such as subspace projection methods, including multiple signal classification (MUSIC)^[15], estimation of signal parameters via rotational invariance techniques (ESPRIT)^[16] and their extensions^[17-18], maximum likelihood methods (MLMs)^[19], iterative adaptive approach for amplitude and phase estimation (IAA-APES)^[20], compressed sensing (CS)^[21-24], and neural network-based approaches^[25]. Overall, these algorithms suffer from high computational complexity to varying degrees. Operations such as matrix inversion and eigenvalue decomposition pose significant challenges for real-time processing in automotive radar systems, especially under cost constraints. Furthermore, the slow convergence speed of IAA-APES and the parameter sensitivity of CS-based algorithms may influence their effectiveness.

For rapidly moving vehicles, the constantly evolving surrounding environment poses challenges to traditional super-resolution algorithms. Moreover, in automotive radar scenarios, targets are often extended targets rather than the ideal point ones encountered in typical DOA estimation scenarios^[26]. This implies that the number of targets is unknown and can change rapidly. Even worse, the equivalent discrete scattering points within an extended target can exhibit significant variations in reflection intensity at any given moment. These characteristics introduce practical challenges for the aforementioned algorithms. For instance, subspace projection methods often require model selection procedures to estimate the number of targets, while the sparsity assumption of CS-based approaches may not hold true in such scenarios^[27].

Therefore, it is imperative to examine how we can empower MMW radar to genuinely cater to the demands of autonomous driving. LiDAR excels in spatial scanning with narrow beams and low sidelobes, enabling precise imaging of target edges and contours. In contrast, MMW radar faces difficulties in

achieving distinct imaging of target contours due to its wider beams and relatively higher sidelobe levels. For example, Fig.1 illustrates a common scenario in real-world automotive radar applications. The radar is mounted on Car 1, and two vehicles travel alongside each other with a specific distance in front. Assuming each vehicle's width is 1.5 m and the gap between them is set to 1 m, the angular values formed by the reflections from the edges of each vehicle are denoted by θ_1 , θ_2 , θ_3 , and θ_4 , respectively.

Given the distance R is 20 m, the angular values are: $\theta_1 = -2.15^\circ$, $\theta_2 = 2.15^\circ$, $\theta_3 = 3.58^\circ$, and $\theta_4 = 7.83^\circ$. Although θ_3 is over 1° larger than θ_2 , distinguishing between a single target and two separate targets using a high-resolution 4D radar with 0.5° angular resolution can be extremely challenging. This is because even a 4D radar of such high resolution is incapable of capturing all scattering points from the target vehicles, only detecting a limited number of points, which subsequently hinders accurate scene comprehension. This also demonstrates the significance of MMW radar in target contour imaging.

Advanced radar systems with higher resolution and denser point clouds are needed to meet the requirements of autonomous driving applications. The enhancement of range resolution has been greatly overlooked in previous studies, despite its potential to help form high-density point cloud output. As is well known, range resolution is inversely proportional to the radar signal bandwidth. To improve range resolution, one direct method is to use a large signal bandwidth. However, directly transmitting large bandwidth signals necessitates a higher sampling rate. Therefore, large bandwidth signals are only used in short-range radar applications, while in medium- or long-range detection radar, the signal bandwidth is typically 100 MHz or 150 MHz, and the corresponding range resolution is 1.5 m or 1 m. If a signal with a 4 GHz bandwidth supported by 77 GHz commercial devices is transmitted, the range resolution can be improved by tens of times, and the number of point clouds formed will also increase by tens of times. Instead of directly transmitting a large bandwidth signal, using frequency agility transmission provides an alternative solution^[28-31].

In addition, TDMA suffers from transmission energy loss due to the fact that only one antenna is active at any moment, and therefore the Doppler division multiple access (DDMA) approach is proposed^[32-33]. Instead of achieving waveform di-

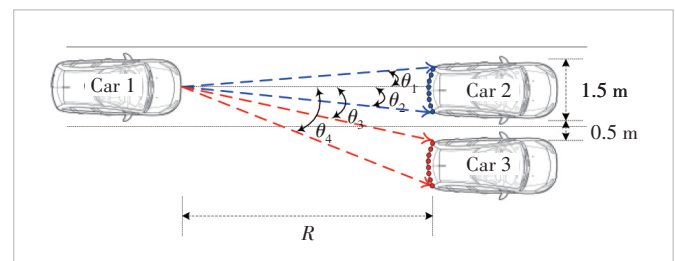


Figure 1. Automotive radar target detection

versity in the time domain, DDMA enables the design of orthogonal waveforms in the Doppler domain and the transmitted waveforms are modulated by multiple sets of linear phase codes. Thus, the transmitted waveforms encoded by DDMA are completely separated in the Doppler domain with various Doppler offsets. An integrated TDMA-DDMA-MIMO framework is proposed in Ref. [34], which combines the advantages of both.

In this paper, a high-resolution automotive radar imaging algorithm that integrates DDMA and frequency agility (FA) is proposed. This approach achieves high resolution in the range dimension, while utilizing DDMA to enhance emission efficiency, enabling high-density point cloud output. Moreover, most of the calculations involved can be completed using FFT, making it highly practical for engineering implementation. The main contributions of this work are summarized as follows:

1) An integrated DDMA-FA-MIMO framework is designed with a simple implementation of orthogonal waveform transmission, which achieves high range resolution with low sampling requirements and high emission efficiency.

2) A detailed analysis is provided regarding the coupling relationship between range and velocity, and a method for simultaneously measuring range and velocity parameters is given, which also takes into account DDMA- and FA-related target detection issues.

3) Based on the frequency-agile and DDMA transmission framework, a detailed analysis is provided on the performance characteristics of clutter echoes in the range-Doppler matrix, as well as the principles of achieving high range resolution.

4) Extensive practical experiments with Texas Instruments (TI) devices are carried out to further confirm and evaluate the effectiveness of the proposed approach.

2 Signal Model

This paper presents a high-resolution approach using a 77 GHz automotive radar system applying a frequency-modulated continuous wave (FMCW) form. We assume a MIMO antenna system composed of N transmitting antennas and L identical receiving antennas, and M pulses will be transmitted in one coherent processing interval (CPI).

In the proposed DDMA-FA-MIMO FMCW radar transmission scheme, an FMCW signal with bandwidth B and duration t_{sub} is transmitted for each pulse, but with a different modulated phase. For the n -th transmitter and m -th pulse, the transmitted signal can be represented as:

$$s_T^n(t, m) = G_n \cdot p_n(m) \cdot \text{rect}\left(\frac{t - t_{\text{sub}}/2}{t_{\text{sub}}}\right) \exp(j2\pi(f_0 + m \cdot \Delta f)t) \cdot \exp(j\pi\mu t^2) \quad (1),$$

where t denotes the time within each pulse, G_n is the gain of the n -th transmitter, f_0 denotes the starting frequency of the first pulse, Δf represents the frequency step between consecu-

tive emitted pulses, t_{sub} is the pulse width, B is the bandwidth of each pulse, m denotes the m -th pulse where $m \in [0, 1, \dots, M-1]$, n denotes the n -th transmitter where $n \in [1, 2, \dots, N]$, and μ represents the chirp rate of linear frequency modulation (LFM) signals. Especially, the vector $p_n(m)$ behaves as a phase modulator related to the pre-defined n -th transmitter to ensure MIMO waveform orthogonality. The signal transmission scheme is depicted in Fig.2.

As shown in Fig. 2, $\varphi_2, \dots, \varphi_N$ are the modulated phases for transmitters and $\varphi_1 = 0$. More generally, for MIMO systems with N transmitters, the phase modulation matrix in discrete form can be expressed as:

$$P = [p_1 \cdots p_n \cdots p_N] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & e^{j\varphi_n} & e^{j\varphi_N} \\ \vdots & \vdots & \vdots \\ 1 & e^{j(N-1)\varphi_n} & e^{j(N-1)\varphi_N} \end{bmatrix} \quad (2).$$

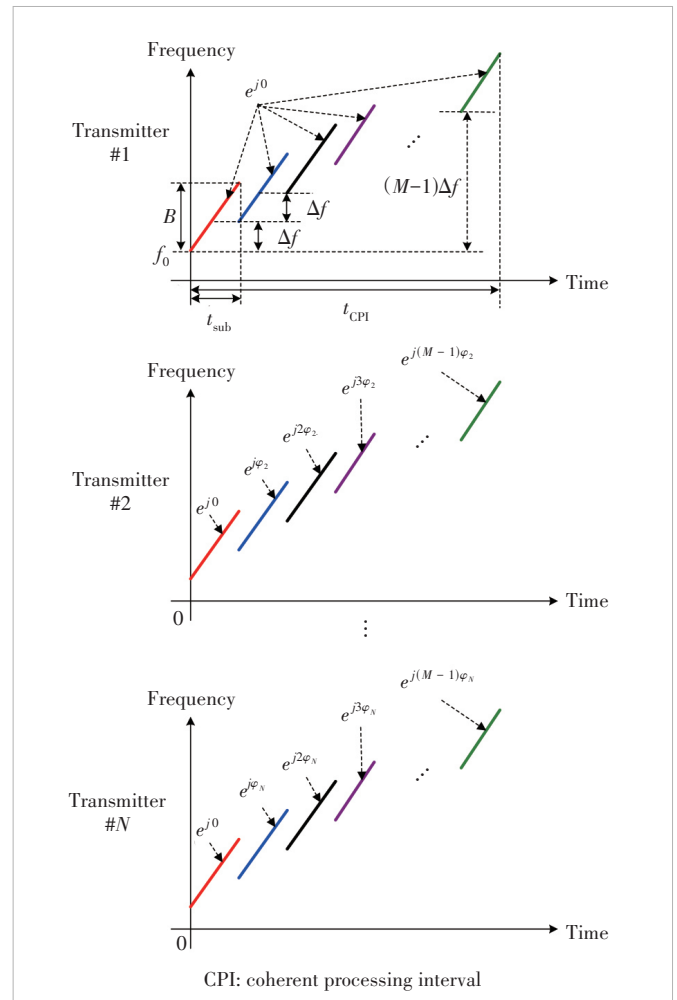


Figure 2. Signal transmission scheme integrating DDMA and FA

For the m -th pulse, the modulated phase vector is the m -th row in matrix $\mathbf{P}$, and the direction of the beam is determined by the currently modulated phase vector. DDMA can be regarded as a time-division phased array radar, and beams are directed towards different directions for different pulses.

In the following, it is assumed that the consecutive sequences do not overlap and that there is no gap between their transmissions. The signal received by the l -th receiver reflected from a moving point-like target located at θ can be written as:

$$s_R^{nl}(t, m, \theta) = \alpha \cdot G_l \cdot s_T^n(t - \tau_{nl}(t), m) \quad (3),$$

where $\tau_{nl}(t)$ denotes the time delay from the n -th transmitting antenna to the target and then to the l -th receiving antenna. The complex factor α includes free space attenuation, radar cross-section (RCS), and the reflection phase of a target. Under the narrowband assumption, when electromagnetic waves impinge on K targets from directions $\theta = [\theta_1, \theta_2, \dots, \theta_K]^T$, the echo can be more concisely expressed as

$$s_R^{nl}(t, m) = G_l \cdot \sum_{k=1}^K \alpha_k \cdot e^{j\varphi_k^{nl}} s_T^n(t - \tau_{0,k}(t), m) \quad (4),$$

where G_l is the gain of l -th receiver, α_k corresponds to the k -th target (also including attenuation, RCS, and reflection phase), and φ_k^{nl} is the delay phase difference caused by the antenna element position. This can be expressed as follows under the far-field assumption:

$$\varphi_k^{nl} = \frac{2\pi}{\lambda} (d_n + d_l) \sin \theta_k \quad (5),$$

where $\lambda = c/(f_0 + m \cdot \Delta f)$ is the wavelength, d_n are the baselines of the n -th transmitter and the reference element, and d_l are the baselines of the l -th receiver and the reference element. For simplicity, a linear array is assumed and only azimuth θ is considered, but the approach is also valid for 3D space when using a 2D array in a 4D radar system.

The round-trip delay $\tau_{0,k}(t)$ for a target with constant velocity can be represented as:

$$\tau_{0,k}(t) = \frac{2(r_{0,k} + v_{0,k}(m \cdot t_{\text{sub}} + t))}{c} = \tau_{0,k} + \frac{2v_{0,k}(m \cdot t_{\text{sub}} + t)}{c} \quad (6),$$

where $r_{0,k}$ denotes the range from the reference transmit antenna (the first transmitter of the radar system) to the k -th target at the beginning of the CPI, $v_{0,k}$ is the velocity, and c is the speed of light.

Generally, MMW radars adopt the dechirp mode in the radio frequency module. By mixing the transmitted and received

signals, the de-chirped intermediate frequency (IF) signal, which will be sampled by an analog-to-digital converter (ADC), can be expressed as:

$$s_{\text{IF}}^{nl}(t, m) = G_l \cdot G_n \cdot p_n(m) \sum_{k=1}^K \alpha_k \cdot e^{j\varphi_k^{nl}} \cdot \exp(-j2\pi(f_0 + m \cdot \Delta f)\tau_{0,k}(t)) \cdot \exp(-j2\pi\mu\tau_{0,k}(t)) \cdot \exp(j\pi\mu\tau_{0,k}^2(t)) \quad (7).$$

Unlike the signal models described in most literature, the frequency variation of Δf between adjacent pulses yields a new phase term $\exp(-j2\pi \cdot m \cdot \Delta f \tau_{0,k}(t))$, which plays a crucial role in achieving high range resolution.

Neglecting the last term and substituting Eq. (6) into Eq. (7), we can get:

$$s_{\text{IF}}^{nl}(t, m) = G_l \cdot G_n \cdot p_n(m) \sum_{k=1}^K \alpha_k \cdot e^{j\varphi_k^{nl}} \cdot \exp(-j2\pi(f_0 + m \cdot \Delta f)\frac{2r_{0,k}}{c}) \cdot \exp(-j2\pi(f_0 + m \cdot \Delta f)\frac{2v_{0,k}(m \cdot t_{\text{sub}} + t)}{c}) \cdot \exp(-j2\pi\mu\frac{2r_{0,k}}{c}) \cdot \exp(-j2\pi\mu\frac{2v_{0,k}(m \cdot t_{\text{sub}} + t)}{c}) \quad (8).$$

Since the Doppler frequency variation caused by different carrier frequencies is small, the Doppler frequency $f_{D,k}$ corresponding to the k -th target is:

$$f_{D,k} = f_0 \frac{2v_{0,k}}{c} = \frac{2v_{0,k}}{\lambda_0} \approx (f_0 + m \cdot \Delta f) \frac{2v_{0,k}}{c} \quad (9).$$

Let $f_{R,k}$ denote the range-dependent frequency component corresponding to the k -th target, which is represented by:

$$f_{R,k} = -\mu \frac{2r_{0,k}}{c} \quad (10).$$

The beat frequency signal denoted by $f_{B,k}$ is

$$f_{B,k} = f_{R,k} + f_{D,k} \quad (11).$$

In automotive radars, the Doppler frequency of the target is typically negligible compared with the frequency resolution of the beat signal, i.e., $f_d = f_s/Q$, where f_s is the sampling frequency of the beat signal and Q is the number of fast-time

samples. Thus, we can approximate $f_{B,k} \approx f_{R,k}$ without loss of generality.

Constant phase terms and some less significant terms can be omitted, leading to:

$$\begin{aligned}
 s_{\text{IF}}^{nl}(t, m) &\approx G_l \cdot G_n \cdot p_n(m) \sum_{k=1}^K \alpha_k \cdot e^{j\varphi_k^{nl}} \cdot \\
 &\exp\left(-j2\pi(m \cdot \Delta f) \frac{2r_{0,k}}{c}\right) \cdot \\
 &\exp\left(-j2\pi \left(\underbrace{f_{D,k} + f_{R,k}}_{f_{B,k}}\right) t\right) \cdot \\
 &\exp\left(-j2\pi f_{D,k} \cdot (m \cdot t_{\text{sub}})\right) = \\
 &G_l \cdot G_n \cdot p_n(m) \sum_{k=1}^K \alpha_k \cdot e^{j\varphi_k^{nl}} \cdot \\
 &\exp\left(-j2\pi f_{R,k} t\right) \cdot \\
 &\exp\left(-j2\pi \left(f_{D,k} t_{\text{sub}} + \frac{2r_{0,k}}{c} \Delta f\right) \cdot m\right)
 \end{aligned} \quad (12).$$

Compared with the traditional signal models in Refs. [7] and [28], the proposed model introduces a new term $\exp\left(-j2\pi(m \cdot \Delta f) \frac{2r_{0,k}}{c}\right)$ related to the frequency agility Δf and the range $r_{0,k}$. Due to this additional term, the Doppler frequency is coupled with range across pulses.

The received signal $s_{\text{IF}}^{nl}(t, m)$ is sampled, and then a fast Fourier transform (FFT) is applied to each individual chirp signal. This splits the radar echo signal into different range gates, and the range resolution is determined by the bandwidth of each pulse B . The beat frequency $f_{B,k}$ can be measured for each target, and the range information of potential targets can be acquired.

For total M pulses, the above procedure is repeated. Then, a second FFT inside each single range gate is applied to the resulting signal spectra. However, it is worth noting that the frequency obtained by the second FFT across pulses is not the Doppler frequency f_D , as it is coupled with the target distance. After the range-dimension FFT (column-wise) and the Doppler-dimension FFT (row-wise), the range-Doppler matrix (RDM) is obtained.

3 Proposed Approach

3.1 Analysis of Range-Doppler Spectra

In traditional waveform diversity MIMO radar systems, the transmitted signals from different transmitters are separated by matched filters, and we can obtain NL RDMs directly. However, for traditional DDMA, the signals are orthogonal in the Doppler domain, so only L RDMs can be obtained, and each RDM will be divided into N matrices. In other words, one tar-

get corresponds to N peaks in one RDM.

For example, in a typical case where two TI AWR2243 chips are cascaded, $N=6$ and $L=8$.

False peaks pose a significant issue in DDMA transmission. The method of transmitting a segment of blank spectrum, as shown in Fig. 3b, is a relatively simple and effective approach. Ref. [32] provides a detailed elaboration on this issue. It should be noted that the frequency $f_{p,n}$ ($n = 1, 2, \dots, N$) obtained by the second FFT across pulses is not the real Doppler frequency f_D of the target, as it also includes the influence of range. Therefore, for targets with the same radial velocity but different ranges, the positions will be different.

3.2 Principles for Obtaining High-Resolution Range-Doppler Spectra

The adopted signal processing flow is illustrated in Fig. 4. For simplicity, a single receiving antenna is taken as an example. After each pulse is received, the same signal is applied to the dechirp process for all receivers, since all transmitted signals are identical only with different modulated phases for DDMA. As analyzed above, for each pulse, DDMA can be viewed as a time-shared phased array radar. Therefore, the received signal is essentially the same, with only differences in amplitude due to the varying gain at the target location, which is caused by the beam pattern formed by the modulated phases between pulses. As a result, reference transmitted signals from one of the transmitting antennas can be used for the entire dechirp process.

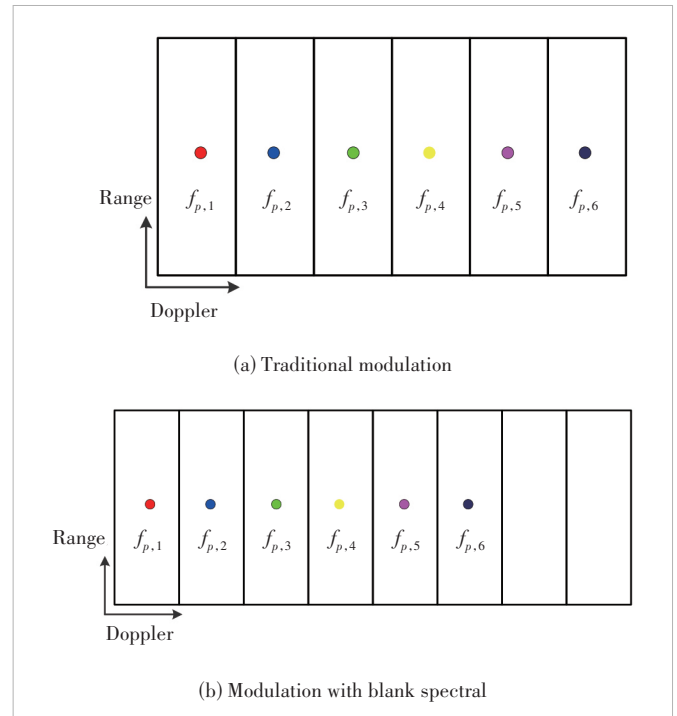


Figure 3. Range-Doppler matrix of a slow-moving target with the traditional DDMA modulation ($N=6$)

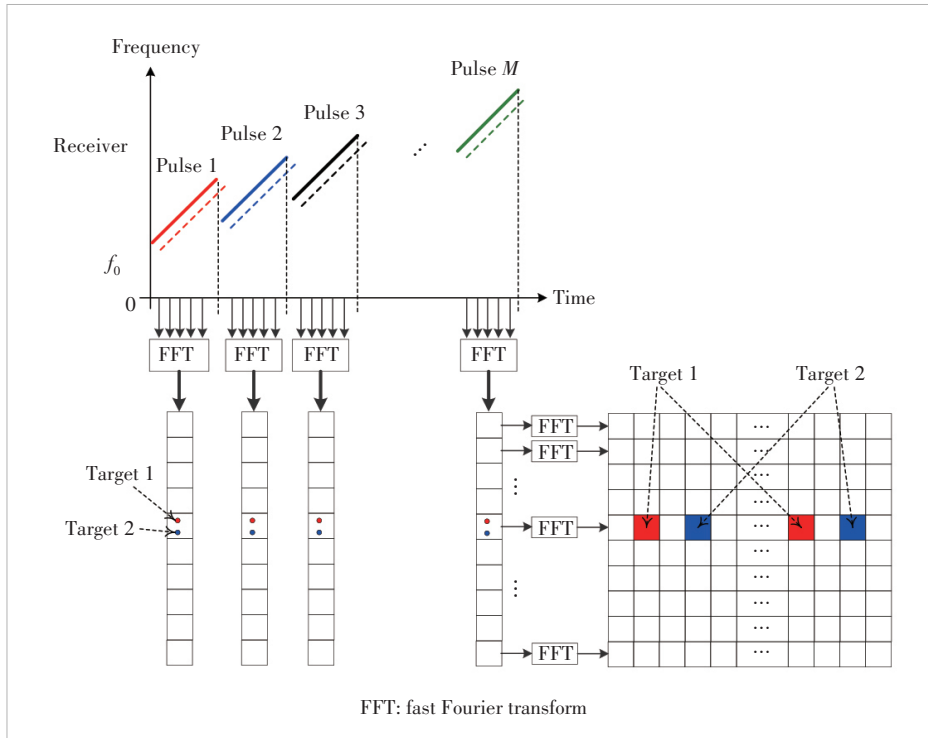


Figure 4. Diagram of signal processing

This paper focuses on improving the range resolution through frequency agility. Two point targets located in the same range cell are configured, as illustrated in Fig. 4. Although two point targets are depicted within a single range cell in the figure, they are indistinguishable in the range-dimension spectrum of each pulse after the dechirping process.

The above derivation also shows that due to the frequency agility across pulses and the distance between the two targets, an additional frequency value can be formed during the FFT processing along the pulse dimension, which is superimposed on the Doppler frequency of the targets. Therefore, if the frequency value resulting from the combined effect of the distance difference and the frequency hopping value is greater than the Doppler resolution, the two point targets will be located within different Doppler cells after performing two-dimensional FFT processing, represented by the red box (target 1) and the blue box (target 2) in Fig. 4. Additionally, due to DDMA

modulation, a target corresponds to multiple peaks, where all the red boxes correspond to target 1.

To illustrate the differences more clearly, the following analysis covers three typical scenarios, including an ideal point target within the range cell ranging from r_1 to $r_1 + \Delta R$, two ideal point targets with slightly different distances, and continuous ground clutter within this range cell. For these three scenarios, we compare the range-Doppler (R-D) spectra obtained using the traditional method of transmitting a chirp pulse train with those of the approach using inter-pulse frequency agility. When there is a difference in the frequency dimension of each transmitted pulse—for instance, when each pulse jumps according to a frequency Δf —the resulting R-D spectrum varies for different radar detection scenarios, as illustrated in Fig. 5.

As shown in the figure, for an ideal stationary point target in the R-D spectrum, the range cell re-

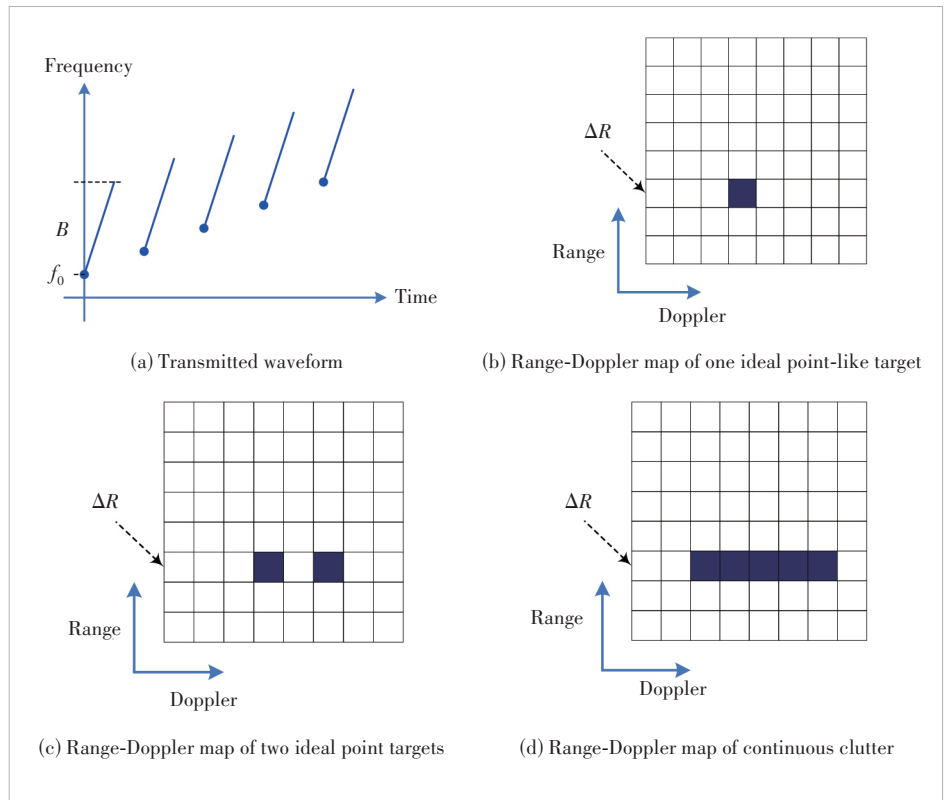


Figure 5. Range-Doppler spectra of typical scenarios for FA FMCW signals

mains consistent with that of a conventional chirp signal, but the Doppler frequency is no longer zero due to DDMA modulation. Similarly, for two ideal point targets, two peaks appear in the R-D spectrum, both exhibiting non-zero Doppler frequencies at different positions in the Doppler dimension. A continuous clutter target corresponds to a continuous velocity interval in Doppler cells.

These effects can be explained by the physical process of waveform propagation. When signals are transmitted at different frequencies, a regular phase change occurs between pulses due to the change in pulse frequency. Performing an FFT on the pulse dimension allows this continuous phase change to manifest as a specific velocity value. Assuming that the precise distance of point target a is r_a , the precise distance of point target b is r_b , and the starting distance of the continuous clutter is r_1 with a distance length of ΔR .

3.3 Proposed Approach for High Resolution Automotive Radar Imaging

This section describes how to obtain a high-resolution R-D spectrum. The process begins with transmitting a sequence of LFM pulses. Then, based on the detected target distance information, the boundaries of the current range bin are determined. Next, the corresponding Doppler channel boundaries are calculated, and a judgment is made on whether Doppler ambiguity exists. After that, a high-resolution range spectrum is obtained, accounting for the presence of Doppler ambiguity. Finally, angle estimation is performed to generate the target contour, and the calculation process can be described as follows:

1) Step 1: Transmit a series of M LFM pulses with a step size of Δf in the frequency hopping sequence. Perform a 2D FFT on the resulting pulse echo data matrix to obtain the R-D spectrum.

2) Step 2: Based on the detected target range bin i , determine the distance interval $[r_{i,\min}, r_{i,\max}]$ for that range segment. The calculation method is as follows:

$$r_{i,\min} = (i - 1) \cdot \Delta R \quad (13),$$

$$r_{i,\max} = i \cdot \Delta R \quad (14),$$

where $\Delta R = c/(2B)$ is the range resolution corresponding to the bandwidth B of a single pulse transmission signal.

3) Step 3: Given the boundaries of the range bin, calculate the corresponding boundaries of the velocity interval and determine the velocity channel number:

$$v_{i,\min} = r_{i,\min} \frac{\Delta f}{f_{\text{start}} \cdot T} \quad (15),$$

$$v_{i,\max} = r_{i,\max} \frac{\Delta f}{f_{\text{start}} \cdot T} \quad (16),$$

where f_{start} is the frequency of the first pulse within the current CPI processed for the velocity dimension FFT, Δf is the frequency step between pulses, and T is the pulse repetition interval.

4) Step 4: Determine the velocity channel number based on the calculated velocity interval boundaries.

Assuming the channel numbers are counted from 0 to $M - 1$, the channel numbers corresponding to the minimum and maximum values of the velocity interval are, respectively:

$$q_{i,\min} = \lfloor v_{i,\min} / \Delta v \rfloor \quad (17),$$

$$q_{i,\max} = \lfloor v_{i,\max} / \Delta v \rfloor \quad (18),$$

where $\lfloor \cdot \rfloor$ denotes the operation of rounding down to the nearest integer.

5) Step 5: Calculate the high-resolution range profile within the velocity interval.

Based on the obtained velocity interval numbers $q_{i,\min}$ and $q_{i,\max}$ corresponding to the current range bin, calculate the output of the high-resolution range profile within that velocity interval:

$$r_{ij} = \frac{(q_{ij} \cdot \Delta v) \cdot f_{\text{start}} \cdot T}{\Delta f} \quad (19),$$

where q_{ij} represents the velocity channel index between $q_{i,\min}$ and $q_{i,\max}$.

6) Step 6: Correct the high-resolution range profile output when resolving velocity ambiguity.

The above formula calculates the distance value r_{ij} , which is valid when there is no velocity ambiguity. If velocity ambiguity occurs, the actual distance value should be corrected to:

$$r_{ij} = \frac{((q_{ij} + (n_{\text{amb}} - 1) \cdot M) \cdot \Delta v) \cdot f_{\text{start}} \cdot T}{\Delta f} \quad (20),$$

where n_{amb} represents the ambiguity order of the velocity channel. When there is no ambiguity, this value is 1. For the current range bin, it can be determined whether the value of the target range exceeds the maximum unambiguous range by checking if:

$$n_{\text{amb}} = \text{floor}(r_i / r_{v,\max}) \quad (21).$$

When the calculated lower and upper boundaries of a range bin differ, it is necessary to split that range bin, i.e.,

$$\text{floor}\left(r_{i,\min}/r_{v,\max}\right) \neq \text{floor}\left(r_{i,\max}/r_{v,\max}\right) \quad (22).$$

When this condition holds, first calculate:

$$n_{\text{amb}} = \text{floor}\left(r_{i,\min}/r_{v,\max}\right) \quad (23).$$

Search for the velocity channels, and when the velocity channel returns from $M - 1$ back to 0, increment the ambiguity order by 1, i.e.,

$$n_{\text{amb}} = n_{\text{amb}} + 1 \quad (24).$$

7) Step 7: Perform DOA estimation on the high-resolution range spectrum.

For each received data matrix, calculate the corresponding high-resolution R-D spectrum matrix. Perform DOA estimation on the first column of this matrix, which corresponds to the zero-velocity channel at different range dimensions. Methods such as DBF or MUSIC can be used for estimation. This allows for an accurate description of clutter distribution in 2D space. Since the presence of different velocities for the same target has a positive effect on target contour imaging, the Doppler spread issue of the target is not specifically addressed in the paper. Instead, it can be processed according to the procedure outlined previously.

4 Simulation and Experimental Results

First, it is verified that high resolution in the range dimension can be achieved through frequency agility between pulses. With a radar frequency f_0 of 77 GHz, a signal bandwidth B of 100 MHz within a single chirp, 1 024 pulses emitted, and a frequency interval Δf of 3 MHz between pulses, the range resolution calculated based on the bandwidth of a single pulse is 1.5 m. Since the actual system verification later in this paper uses a cascade of two TI AWR2243 commercial chips, to ensure consistency with the actual radar system, the simulation here also adopts a 6-transmitter 8-receiver MIMO array, which is equivalent to a uniform linear array containing 48 elements arranged at half-wavelength intervals.

Assume that there are three ideal point targets located at 150.2 m, 150.5 m, and 151 m, respectively. If processed using the traditional non-frequency agile pulse train mode, these three targets would all fall within the same range cell and could not be resolved. However, by adopting a frequency agile pulse emission mode, the range resolution after synthesizing a wideband signal is approximately 0.049 m, allowing for the complete differentiation of the three targets.

In the following simulations, to closely approximate real-world testing scenarios, the simulated objects are modeled as extended targets, such as vehicles 5 m in length and 1.5 m in width. These models contain a large number of randomly distributed scattering points, with random positions, reflection in-

tensities, and phases. The performance of different algorithms is compared. Traditional angle measurement methods using pulse trains with the same carrier frequency employ the MUSIC algorithm. However, for the DDMA-FA-MIMO-based algorithm, only the simplest DBF method is considered for angle measurement. In practical applications, if the computational complexity is acceptable, some super-resolution algorithms can also be integrated into this processing framework. The range-Doppler spectrum for a single vehicle is shown in Fig. 6.

In Fig. 7, the DDMA-FA-MIMO method generates denser and more concentrated radar point clouds than the traditional method, with the colorbar indicating signal reflection intensity. Based on the point cloud output results, the outlines of the vehicles can be accurately determined.

We conducted real-world verification using a radar system consisting of two cascaded TI AWR2243 commercial chips. The radar detection results for real-world road scenarios are

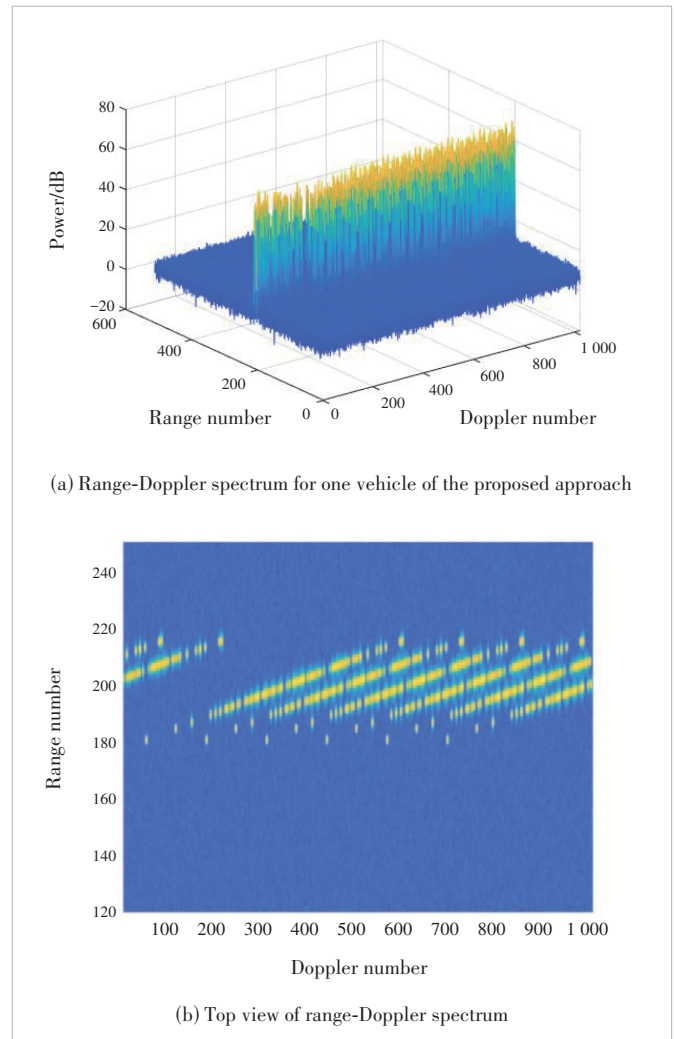


Figure 6. Range-Doppler spectrum and top-view imaging of the proposed MIMO radar approach

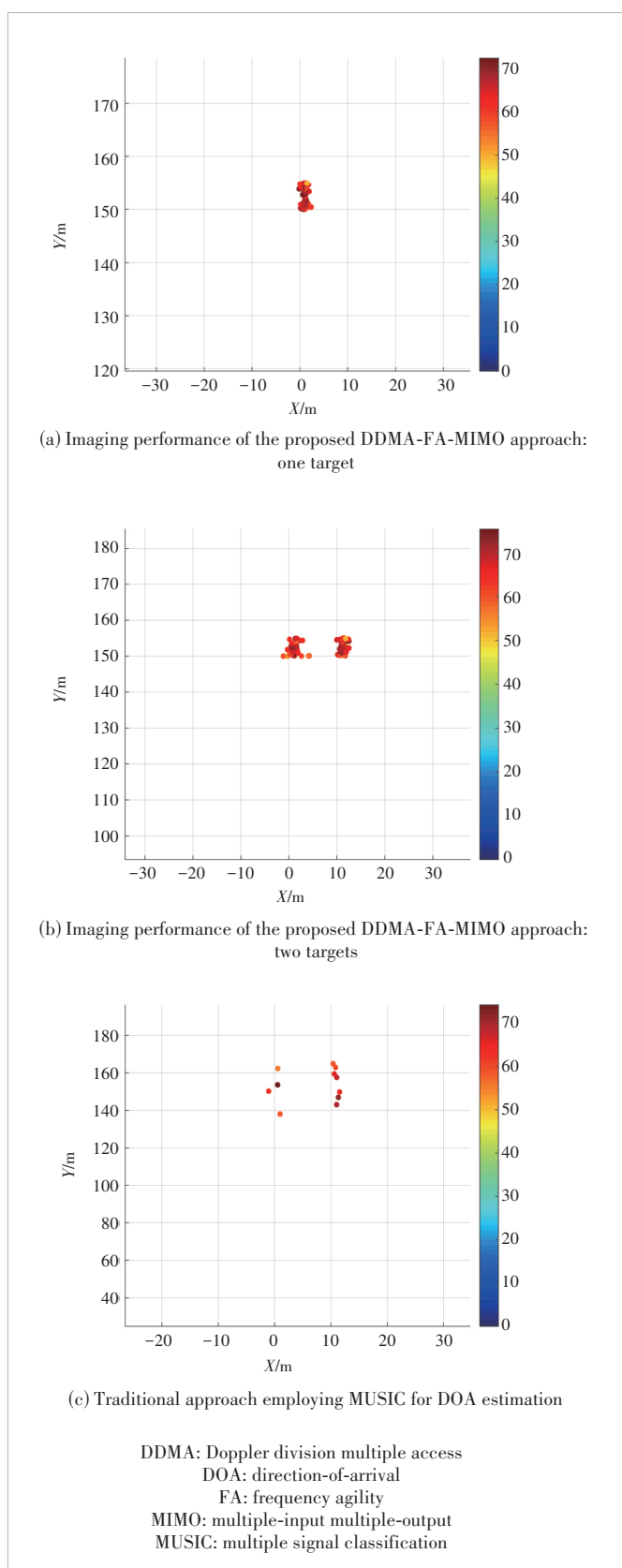


Figure 7. Comparison of target imaging performance between traditional method and DDMA-FA-MIMO approach

presented in Fig. 8, where stationary pedestrians and vehicles are clearly identified, and the imaging of various vehicles and pedestrians at near and far distances is all very clear.

To avoid vehicle obstructions encountered during ground-level installation, the radar was deployed on an overhead bridge. Separate tests were performed for oncoming and receding vehicle targets, as shown in Figs. 8a and 8b. The results reveal highly dense point-cloud outputs with well-defined target outlines. Although road surface reflections are not ideal specular reflections, they are relatively continuous. CFAR processing was applied to detect target contour output^[35–36], which effectively removed sporadic noise points. Therefore, grass or road edges were clearly detected, while there was no detection output for the road surface. Both moving buses and stationary roadside vehicles have distinct contours. Fig. 8c depicts the ground-based radar test scenario. The size of the point cloud envelope is related to the target's height, and vehicle contours are significantly larger than those of motorcycles. These contours enable clear identification of various road targets.

5 Conclusions

The high angular resolution of LiDAR enables it to play a crucial role in ADAS and even autonomous driving. To enhance the competitive advantage of MMW radar over LiDAR, extensive research has been conducted on improving angular resolution. However, it is worth noting that LiDAR achieves its high angular resolution through a very narrow emission beam. If MMW radar could generate a very narrow emission beam while maintaining low sidelobes, it could achieve similar effects as LiDAR does. Nevertheless, achieving such a narrow emission beam comes with significant costs. Currently, MMW radar relies primarily on backend angle super-resolution, compressed sensing, and intelligent algorithms to achieve high angular resolution.

In practical automotive radar applications, targets are typically extended targets rather than isolated ideal point targets. This results in many equivalent scattering points, significant differences in the strengths of these scattering points, and rapid dynamic changes in equivalent scattering points. These factors limit the performance of the aforementioned algorithms in practical applications.

Based on the analysis of practical automotive radar application scenarios, this paper highlights the significant value of range resolution, suggesting that the traditional challenge of angular discrimination can be transformed into a range discrimination problem. Since range resolution is directly related to the bandwidth of the transmitted signal, and transmitting a large signal bandwidth for medium- to long-range MMW radars significantly enhances the sampling rate, which increases costs and is unachievable with current commercial devices. Therefore, this paper proposes a method of pulse-dimension frequency agility to synthesize wideband signals.

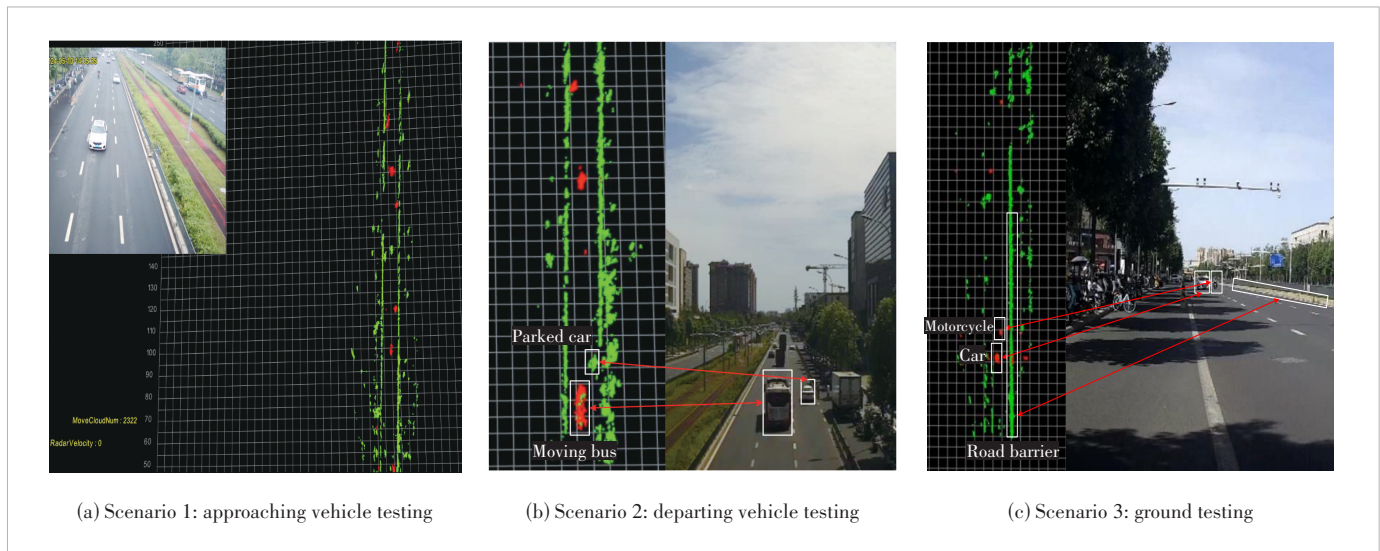


Figure 8. Field test results of practical radar system

At the same time, to increase signal transmission efficiency, this paper combines the frequency-agile transmission method with the DDMA orthogonal MIMO transmission method, proposing the DDMA-FA-MIMO transmission framework. It conducts a detailed analysis of scenarios involving ideal single-point and multi-point targets, as well as continuous clutter targets, under this framework. The paper elucidates the physical mechanism of high-resolution imaging and provides the implementation process of the algorithm.

Through theoretical simulations and practical system tests using the dual-cascaded TI AWR2243 commercial chip, the proposed approach has been verified to generate high-density point clouds for targets and realize preliminary imaging of target contours. These capabilities provide a solid foundation for target recognition in ADAS and autonomous driving systems.

References

- [1] Patole S M, Torlak M, Wang D, et al. Automotive radars: a review of signal processing techniques [J]. *IEEE signal processing magazine*, 2017, 34(2): 22 – 35. DOI: 10.1109/msp.2016.2628914
- [2] Hakobyan G, Yang B. High-performance automotive radar: a review of signal processing algorithms and modulation schemes [J]. *IEEE signal processing magazine*, 2019, 36(5): 32 – 44. DOI: 10.1109/msp.2019.2911722
- [3] Engels F, Heidenreich P, Wintermantel M, et al. Automotive radar signal processing: research directions and practical challenges [J]. *IEEE journal of selected topics in signal processing*, 2021, 15(4): 865 – 878. DOI: 10.1109/jstsp.2021.3063666
- [4] An analysis of Tesla's readoption of millimeter-wave radar [EB/OL]. (2022-06-27) [2024-10-20]. https://www.sohu.com/a/561258382_121124379
- [5] Ge S G, Li K, Rum S N B M. Deep learning approach in DOA estimation: a systematic literature review [J]. *Mobile information systems*, 2021: 6392875. DOI: 10.1155/2021/6392875
- [6] Bechter J, Roos F, Waldschmidt C. Compensation of motion-induced phase errors in TDM MIMO radars [J]. *IEEE microwave and wireless components letters*, 2017, 27(12): 1164 – 1166. DOI: 10.1109/lmwc.2017.2751301
- [7] Sammartino P F, Baker C J, Griffiths H D. Frequency diverse MIMO techniques for radar [J]. *IEEE transactions on aerospace and electronic systems*, 2013, 49(1): 201 – 222. DOI: 10.1109/taes.2013.6404099
- [8] Pfeffer C, Feger R, Wagner C, et al. FMCW MIMO radar system for frequency-division multiple TX-beamforming [J]. *IEEE transactions on microwave theory and techniques*, 2013, 61(12): 4262 – 4274. DOI: 10.1109/tmtt.2013.2287675
- [9] Texas instruments. 76-GHz to 81-GHz automotive second-generation high-performance MMIC [EB/OL]. [2024-10-20]. <https://www.ti.com/product/AWR2243>
- [10] NXP. NXP semiconductors [EB/OL]. [2024-02-14]. <https://www.nxp.com.cn/products/PCPRODCAT>
- [11] Continental Automotive. Continental earns three CES 2021 Innovation Honoree Awards [EB/OL]. [2024-10-20]. <https://www.continental.com/en/press/press-releases/ces-2021-innovation-awards/>
- [12] Iqbal H, Sajjad M B, Mueller M, et al. SAR imaging in an automotive scenario [C]//The 15th Mediterranean Microwave Symposium (MMS). IEEE, 2015: 1 – 4. DOI: 10.1109/mms.2015.7375430
- [13] Zhang W, Wang P, He N Y, et al. Super resolution DOA based on relative motion for FMCW automotive radar [J]. *IEEE transactions on vehicular technology*, 2020, 69(8): 8698 – 8709. DOI: 10.1109/tvt.2020.2999640
- [14] Gao X Y, Roy S, Xing G B. MIMO-SAR: a hierarchical high-resolution imaging algorithm for mmWave FMCW radar in autonomous driving [J]. *IEEE transactions on vehicular technology*, 2021, 70(8): 7322 – 7334. DOI: 10.1109/tvt.2021.3092355
- [15] Schmidt R. Multiple emitter location and signal parameter estimation [J]. *IEEE transactions on antennas and propagation*, 1986, 34(3): 276 – 280. DOI: 10.1109/tap.1986.1143830
- [16] Roy R, Kailath T. ESPRIT-estimation of signal parameters via rotational invariance techniques [J]. *IEEE transactions on acoustics, speech, and signal processing*, 1989, 37(7): 984 – 995. DOI: 10.1109/29.32276
- [17] Stoica P, Wang Z S, Li J. Extended derivations of MUSIC in the presence of steering vector errors [J]. *IEEE transactions on signal processing*, 2005, 53(3): 1209 – 1211. DOI: 10.1109/tsp.2004.842201
- [18] Li J. Improving ESPRIT via beamforming [J]. *IEEE transactions on aerospace and electronic systems*, 1992, 28(2): 520 – 528. DOI:

- 10.1109/7.144577
- [19] Bresler Y, Macovski A. Exact maximum likelihood estimation of superimposed exponential signals in noise [C]/International Conference on Acoustics, Speech, and Signal Processing. Tampa: IEEE, 1985: 1824 – 1827. DOI: 10.1109/ICASSP.1985.1168514
- [20] Yardibi T, Li J, Stoica P, et al. Source localization and sensing: a non-parametric iterative adaptive approach based on weighted least squares [J]. IEEE transactions on aerospace and electronic systems, 2010, 46(1): 425 – 443. DOI: 10.1109/taes.2010.5417172
- [21] Rebollo-Neira L, Lowe D. Optimized orthogonal matching pursuit approach [J]. IEEE signal processing letters, 2002, 9(4): 137 – 140. DOI: 10.1109/lsp.2002.1001652
- [22] Aguilera E, Nannini M, Reigber A. Wavelet-based compressed sensing for SAR tomography of forested areas [J]. IEEE transactions on geoscience and remote sensing, 2013, 51(12): 5283 – 5295. DOI: 10.1109/tgrs.2012.2231081
- [23] Xu G, Pei H, Jiang M J, et al. High-resolution mmWave SAR imagery for automotive parking assistance [J]. IEEE journal on miniaturization for air and space systems, 2023, 4(1): 54 – 61. DOI: 10.1109/jmass.2022.3226771
- [24] Malioutov D, Cetin M, Willsky A S. A sparse signal reconstruction perspective for source localization with sensor arrays [J]. IEEE transactions on signal processing, 2005, 53(8): 3010 – 3022. DOI: 10.1109/tsp.2005.850882
- [25] Fuchs J, Weigel R, Gardill M. Single-snapshot direction-of-arrival estimation of multiple targets using a multi-layer perceptron [C]/International Conference on Microwaves for Intelligent Mobility (ICMIM). IEEE, 2019: 1 – 4. DOI: 10.1109/icmim.2019.8726554
- [26] Buddendick H, Eibert T F. Incoherent scattering-center representations and parameterizations for automobiles [EM Programmer's Notebook] [J]. IEEE antennas and propagation magazine, 2012, 54(1): 140 – 148. DOI: 10.1109/map.2012.6202527
- [27] Roldan I, Fioranelli F, Yarovoy A. Total variation compressive sensing for 3D shape estimation in short-range imaging radars [J]. IEEE transactions on radar systems, 2023, 1: 583 – 592. DOI: 10.1109/trs.2023.3322630
- [28] Garvin A D M. Range, resolution and imaging applications of a stepped-frequency continuous wave radar [J]. Transactions of the South African institute of electrical engineers, 1991, 82(3): 206 – 216
- [29] Huang P H, Dong S S, Liu X Z, et al. A coherent integration method for moving target detection using frequency agile radar [J]. IEEE geoscience and remote sensing letters, 2019, 16(2): 206 – 210. DOI: 10.1109/lgrs.2018.2870869
- [30] Huang T Y, Liu Y M, Ma D Y, et al. Multi-carrier agile phased array radar [C]/International Conference on Signal, Information and Data Processing (ICSIDP). IEEE, 2019: 1 – 6. DOI: 10.1109/icsidp47821.2019.9173350
- [31] Winkler V. Novel waveform generation principle for short-range FMCW-radars [C]/German Microwave Conference. IEEE, 2009: 1 – 4. DOI: 10.1109/gemic.2009.4815876
- [32] Xu F, Vorobyov S A, Yang F W. Transmit beamspace DDMA based automotive MIMO radar [J]. IEEE transactions on vehicular technology, 2022, 71(2): 1669 – 1684. DOI: 10.1109/tvt.2021.3132886
- [33] Bialer O, Jonas A, Turer T. Code optimization for fast chirp FMCW automotive MIMO radar [J]. IEEE transactions on vehicular technology, 2021, 70(8): 7582 – 7593. DOI: 10.1109/tvt.2021.3095956
- [34] Jiang M J, Xu G, Pei H, et al. 4D high-resolution imagery of point clouds for automotive mmWave radar [J]. IEEE transactions on intelligent transportation systems, 2024, 25(1): 998 – 1012. DOI: 10.1109/tits.2023.3258688
- [35] Zhang W, Li H Y, Sun G H, et al. Enhanced detection of Doppler-spread targets for FMCW radar [J]. IEEE transactions on aerospace and electronic systems, 2019, 55(4): 2066 – 2078. DOI: 10.1109/taes.2019.2925433
- [36] Rohling H. Radar CFAR thresholding in clutter and multiple target situations [J]. IEEE transactions on aerospace and electronic systems, 1983, 19(4): 608 – 621. DOI: 10.1109/taes.1983.309350

Biographies

Zhang Wei (zhw5282411@uestc.edu.cn) received his BS, MS, and PhD degrees in electrical engineering from Southwest Jiaotong University, China and the University of Electronic Science and Technology of China (UESTC), in 2004, 2008, and 2013, respectively. He is currently a professor with UESTC. His research interests mainly include array signal processing, millimeter-wave radar, automotive radar, space-time adaptive processing (STAP), and MIMO radar signal processing.

Zhang Chenyong received his MS degree in electronics and communication engineering from the University of Electronic Science and Technology of China. He is currently a radar signal processing engineer. His research interests include the research and development of millimeter-wave radar and phased array radar systems. His main work involves radar target detection, space-time adaptive processing, and FPGA real-time implementation for radar signal processing.

Wang Yu received his bachelor's and master's degrees from the University of Electronic Science and Technology of China in 2013 and 2016, respectively. He has been engaged in theoretical research and engineering implementation of millimeter-wave radar detection and imaging technology. His current research interests mainly include millimeter-wave radar signal processing, high-precision detection and imaging algorithms, and their related engineering applications.

Ding Jin received her BS degree in electronic information from the School of Physics in 2022 and MS degree in communication engineering from the School of Information and Communication Engineering in 2025, both from the University of Electronic Science and Technology of China. Her research interests include automotive 4D millimeter-wave radar, point cloud and image fusion.

Li Yao received his MS degree from the University of Electronic Science and Technology of China. He is now an engineer with Chengdu Nanoradar Technology Co., Ltd. His research interests cover millimeter-wave radar signal processing and point cloud processing.



Constructing Large-Scale Real-World Benchmark Datasets for AIOps

Li Zeyan¹, Han Jing², Zhao Nengwen¹,
Zhang Shenglin³, Sun Yongqian³, Chen Pengfei⁴,
Wen Xidao¹, Ma Minghua⁵, Pei Dan¹

(1. Tsinghua University, Beijing 100084, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. Nankai University, Tianjin 300071, China;
4. Sun Yat-sen University, Guangzhou 510275, China;
5. Microsoft Research, Redmond, WA 98052, USA)

DOI: 10.12142/ZTECOM.202603008

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260903.1213.002.html>,
published online September 3, 2026

Manuscript received: 2025-02-21

Abstract: Artificial Intelligence for IT Operations (AIOps) has emerged as a critical approach for managing complex IT systems, yet its advancement has been hindered by the lack of comprehensive public datasets for evaluation. This study addresses this challenge by introducing three large-scale, real-world benchmark datasets specifically designed for AIOps research and development. We present datasets covering key AIOps scenarios: key performance indicator (KPI) anomaly detection with diverse patterns from multiple Internet companies, multi-dimensional root cause localization with 400 labeled cases, and failure discovery and diagnosis containing 169 injected failures across various system components. These datasets are distinguished by their scale, authenticity, and detailed ground-truth labels, enabling rigorous evaluation of AIOps algorithms. The practical value of our datasets has been validated through three annual AIOps algorithm competitions (2018–2020), which attracted over 400 participating teams and led to numerous research publications. By providing these comprehensive benchmarks, our work establishes a foundation for reproducible AIOps research and facilitates fair comparison of different approaches, serving a role for AIOps analogous to that of ImageNet in computer vision.

Keywords: AIOps; benchmark datasets; KPI anomaly detection; root cause localization; failure discovery and diagnosis

Citation (Format 1): Li Z Y, Han J, Zhao N W, et. al. Constructing large-scale real-world benchmark datasets for AIOps [J]. ZTE Communications, 2026, 24(3): 59 – 69. DOI: 10.12142/ZTECOM.202603008

Citation (Format 2): Z. Y. Li, J. Han, N. W. Zhao, et. al., “Constructing large-scale real-world benchmark datasets for AIOps,” *ZTE Communications*, vol. 24, no. 3, pp. 59 – 69, Sept. 2026. doi: 10.12142/ZTECOM.202603008.

1 Introduction

In recent years, the growing complexity of modern IT service architectures has made traditional manual operations and maintenance increasingly time-consuming and prone to errors. To address this challenge, the concept of Artificial Intelligence for IT Operations (AIOps) has emerged to reduce manual intervention. According to Gartner^[1], AIOps refers to the utilization of machine learning (ML) and data mining techniques in IT operations. By combining the power of big data and ML, AIOps aims to enhance and, to some extent, replace conventional manual tasks such as performance monitoring, anomaly detection, event correlation and analysis, and incident management.

Extensive research efforts have been dedicated to various

aspects of AIOps in the literature, such as anomaly detection^[2–5], root cause localization^[6–17], failure prediction^[18–23], incident management^[24–27], and resource management^[28–29]. However, a major challenge arises from the absence of publicly available large-scale real-world datasets, which makes it difficult to conduct fair comparisons of these approaches.

While these works demonstrate promising performance on their respective private datasets, their real-world robustness remains uncertain. Moreover, the diversity of target AIOps systems makes it challenging to generalize approaches to real-world applications. Consequently, AIOps represents a convergence point between artificial intelligence research and operations engineering, necessitating collaboration between academia and industry to access authentic data and scenarios for research purposes.

In practice, researchers often face challenges in finding industrial partners due to the highly confidential nature of the monitoring data used in AIOps. Consequently, the lack of

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20231009003.

public large-scale real-world datasets remains a significant limitation in both AIOps research and practical application.

While there are a few public datasets available for AIOps research, these datasets suffer from several limitations. First, most existing datasets focus on a single AIOps scenario, such as Yahoo S5^[30] for key performance indicator (KPI) anomaly detection and the system-trace corpora generated from the Sock Shop and TrainTicket microservice benchmarks to evaluate MEPFL^[31]. However, AIOps encompasses a wide range of scenarios, including KPI anomaly detection, log anomaly detection, capacity planning, root cause localization, and more. Evaluating multi-scenario AIOps approaches on scenario-specific datasets individually is inadequate to capture the holistic nature of AIOps. Second, the scale of these datasets is limited. For instance, Yahoo Benchmark for KPI anomaly detection consists of only hundreds of data points. In reality, service systems often support numerous users and operate in complex architectures. Therefore, to approximate real-world conditions, datasets should be large and diverse. Third, some datasets are synthetic rather than real-world. Such synthetic datasets fail to accurately reflect the behavior of actual systems, which poses risks for deploying AIOps techniques in practice. Consequently, there is a pressing need for rich, large-scale, and real-world datasets to facilitate the advancement of AIOps research.

To foster the development of AIOps, we have released three large-scale, real-world datasets on GitHub^[32–34]. These datasets are sourced from the production systems of our industrial partners, including Sogou, eBay, Tencent, Suning, and China Mobile Zhejiang. Each dataset caters to a specific aspect of AIOps: KPI anomaly detection, multi-dimensional root cause localization, and comprehensive failure discovery and diagnosis. Numerous studies^[6, 10, 35–41] have already utilized these datasets. Furthermore, based on these datasets, we have organized an annual AIOps algorithm competition since 2018. We organized three AIOps algorithm competitions in 2018, 2019, and 2020. These editions attracted 125, 141, and 141 teams, respectively, totaling 407 team registrations and more than 1 000 participant registrations.¹ To encourage collaboration and knowledge sharing, we have publicly released the solutions from the top teams^[42]. Moving forward, we remain committed to creating and publishing new datasets, as well as organizing the AIOps algorithm competition annually.

The main contributions of our work are summarized as follows.

- **Publication of large-scale real-world datasets:** We have released three large-scale real-world datasets^[32–34], covering essential aspects of AIOps such as KPI anomaly detection, root cause localization for multi-dimensional data, and fail-

ure discovery and diagnosis. These datasets have served as valuable resources and have been utilized by many research works in the field.

- **Systematic dataset design and validation:** Our datasets are constructed through a well-defined methodology ensuring data quality and comprehensive coverage of real-world scenarios. We have implemented a rigorous failure injection process and provided expert labeling from domain practitioners. The practical value of these datasets has been validated through their use in numerous studies and annual AIOps algorithm competitions.

2 Our Work

Building upon existing work^[43], we categorize AIOps research into two main areas: failure management and resource provisioning, as illustrated in Fig. 1. Failure management focuses on minimizing the negative impact of failures, with the ultimate goal of ensuring service availability and optimal user experience. This area encompasses various techniques and approaches to mitigate failure impact. Meanwhile, resource provisioning pertains to the allocation and optimization of diverse resources, such as power and virtual machines (VMs), in the context of IT service management. This aspect of AIOps aims to effectively manage and utilize resources to meet service demands while optimizing resource consumption and costs. At present, our primary focus lies within the domain of failure management. However, we acknowledge that resource provisioning is an important aspect that warrants attention and could potentially be a focus area for our

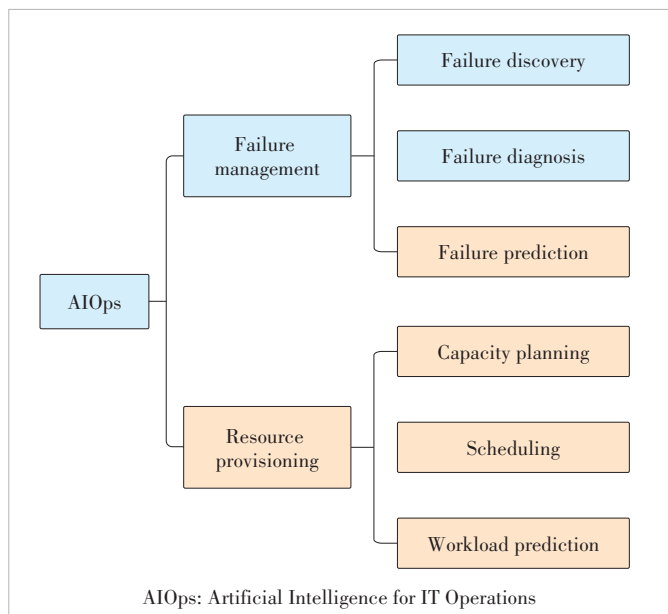


Figure 1. Taxonomy of AIOps research

¹ <https://competition.aiops-challenge.com/home/competition>

future work.

As depicted in Fig. 2, mean time to detect (MTTD) measures failure-discovery latency, mean time to restore (MTTR) measures the time required to diagnose and restore services after a failure, and mean time between failures (MTBF) measures service reliability between successive failures. To address these limitations, we draw inspiration from ImageNet^[44] and aim to develop large-scale real-world benchmarks for AIOps. Our initial focus is on failure management, and we have published three related datasets. These datasets encompass KPI anomaly detection (Section 3), root cause localization for multi-dimensional data (Section 4), and failure discovery and diagnosis (Section 5). In the subsequent sections, we provide an overview of each dataset and its background.

3 KPI Anomaly Detection

In this section, we provide an overview of the background of KPI anomaly detection, followed by a detailed description of the dataset we have published to facilitate research in this area.

3.1 Background

To monitor the operational status of services and promptly identify system failures, engineers often collect a multitude of KPIs and analyze them to detect anomalies. Accurate KPI anomaly detection plays a crucial role in triggering timely troubleshooting, minimizing economic losses, and maintaining a satisfactory user experience^[3-4, 36, 45].

Over the years, substantial efforts have been devoted to KPI anomaly detection. Existing approaches can be broadly categorized into three main types: supervised methods (e.g., Opprentice^[4] and EGADS^[30]), unsupervised methods (e.g., Donut^[3]), and semi-supervised methods (e.g., ADS^[2]). Each method offers unique techniques for KPI anomaly detection.

Despite these advancements, several significant challenges still impede research progress in KPI anomaly detection^[36]. These challenges need to be addressed to enhance the effectiveness and efficiency of anomaly detection methods in real-world scenarios.

- **Various KPI patterns:** KPIs exhibit diverse patterns, which can be classified into three main categories: seasonal (e.g., transaction volume), stable (e.g., success rate), and un-

stable (e.g., CPU utilization). A successful KPI anomaly detection approach should generalize well across different pattern types, ensuring its applicability in various scenarios.

- **Diverse KPI anomaly patterns:** KPI anomalies can manifest in different forms, such as spikes, dips, and jitters, representing deviations from normal behavior. Moreover, different scenarios and components may have varying tolerance levels towards anomalies. Therefore, an effective approach should be capable of detecting diverse anomaly patterns and adapting to different contexts.

- **Limited availability of labeled data:** Obtaining accurately labeled KPI data is challenging. Large-scale systems often generate a vast number of KPIs, making it difficult to label anomalies manually. Moreover, labeling anomalies requires domain experts, further increasing the time and effort required. Consequently, the scarcity of labeled data hampers the construction of supervised models and the evaluation of detection performance.

Despite the existence of numerous KPI anomaly detection approaches, their evaluation is often limited to private datasets, and the lack of large-scale real-world datasets hinders their practical deployment in industry. Public benchmarks, such as Yahoo S5^[30] and the Numenta Anomaly Benchmark^[46], contain mixtures of real-world and synthetic or artificial series and have substantially shorter series than Dataset A.

In reality, the KPIs in large-scale systems are numerous and diverse. Therefore, there is a pressing need for a comprehensive, large-scale, and diverse KPI dataset for anomaly detection. Such a dataset would enable researchers and practitioners to develop and evaluate their approaches on realistic scenarios, bridging the gap between promising research results and real-world performance.

By addressing these limitations and providing a rich dataset for KPI anomaly detection, we aim to facilitate the development of more effective and practical anomaly detection techniques for real-world deployment.

3.2 Dataset

To address the aforementioned challenges, we have developed a large-scale real-world KPI anomaly detection dataset, referred to as Dataset A. This dataset has been constructed to encompass diverse KPI and anomaly patterns, and it provides sufficient labeled data for model training and evaluation. Dataset A was collected from Sogou, eBay, Baidu, Tencent, and Alibaba; Dataset B was derived from production data at Suning; and Dataset C was collected from a production system at China Mobile Zhejiang.

Dataset A consists of 7 KPIs, each of which has been labeled for anomalies by experienced engineers from these five companies. This ensures the accuracy and reliability of the labeled data. In Table 1, we provide some key statistics about Dataset A, including the total number of data points, anomaly ratio, and time span. Notably, the time span of each

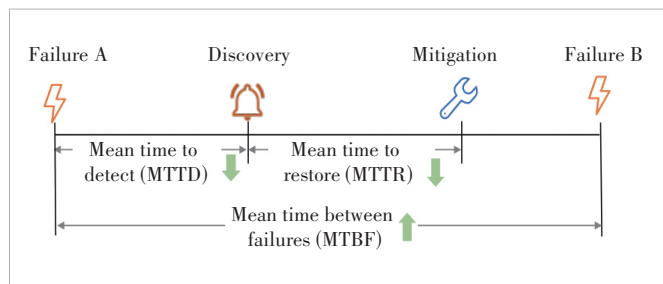


Figure 2. Pipeline and key measurements in failure management

KPI in the dataset ranges from two months to seven months, which demonstrates its large-scale nature compared with existing datasets like Yahoo^[30]. Fig. 3 illustrates five selected KPIs within a one-week time span from Dataset A, reflecting the diverse nature of the dataset.

By providing such a comprehensive dataset, we aim to facilitate research and development in KPI anomaly detection, enabling the exploration and evaluation of novel approaches in real-world scenarios.

Utilizing Dataset A, we organized the inaugural AIOps algorithm challenge in 2018, which garnered significant participation from both academia and industry. The challenge attracted 125 teams and over 300 participants, highlighting the strong interest in the AIOps community. Detailed information about the competition can be found on the competition website.²

To evaluate the performance of KPI anomaly detection algorithms effectively, we introduced a novel evaluation method inspired by the alarm delay restriction proposed in Donut^[3]. This evaluation method, which takes into account alarm delay constraints, was specifically designed for the competition. To facilitate further research and development, we have made the evaluation method available on GitHub,³ allowing researchers to assess the performance of their algorithms using the same metrics and constraints.

Since its release, Dataset A and the associated evaluation method have been widely adopted and referenced in various research works, including papers such as Refs. [35 – 37, 41]. This demonstrates the utility and impact of the dataset and evaluation method for KPI anomaly detection.

During the competition, the best F1-score achieved by the participating teams reached 0.821 6, showcasing the high-quality solutions for KPI anomaly detection. Moreover, the solutions proposed by the top teams have been publicly shared,⁴ allowing for further analysis and collaboration among researchers and practitioners.

As shown in Table 2, Dataset A demonstrates several advantages compared to existing benchmarks in KPI anomaly

Table 1. Summary of the preliminary-stage release of Dataset A

Statistic	Value	Range	Mean
KPI series	26	-	-
Data points	4 821 526	16 495 – 295 379	185 443.31
Reported anomaly ratio	-	0.01% – 4.37%	0.96%
Time span/d	-	60 – 207	148.31

KPI: key performance indicator

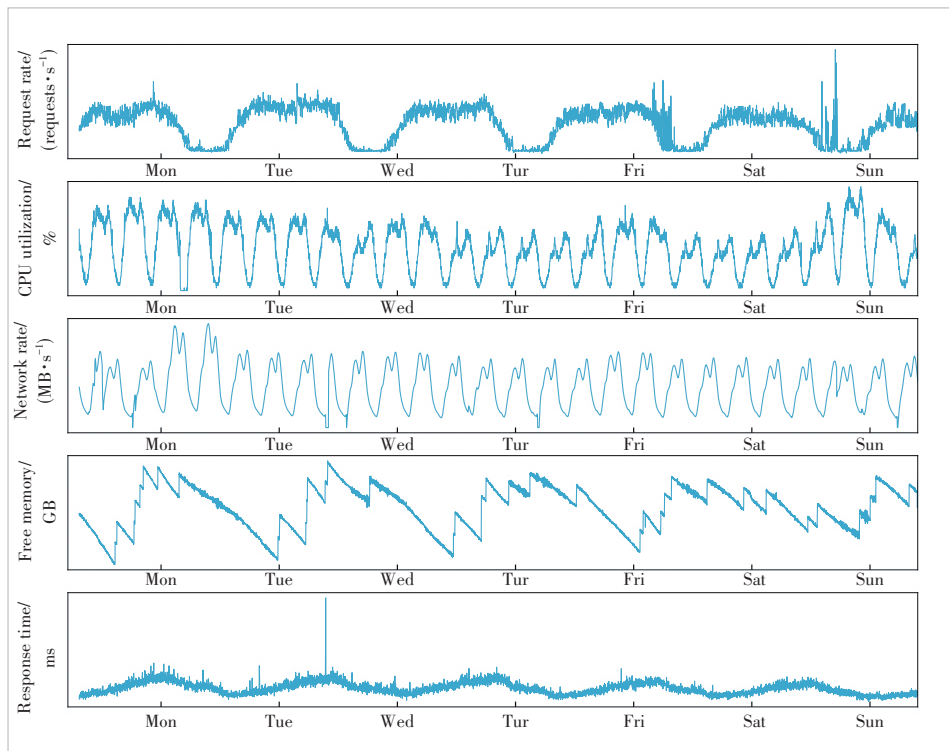


Figure 3. Some KPI examples of Dataset A

detection. With an average length of 211 969.75 data points, Dataset A is significantly larger than other widely-used datasets such as WSD (36 471.00) and UCR (55 935.25), providing a more robust foundation for large-scale evaluation. Furthermore, like other datasets in our comparison, Dataset A originates from real-world data, ensuring representative evaluation under real-world conditions.

4 Root-Cause Localization for Multi-Dimensional Data

In this section, we first introduce the background of root cause localization for multi-dimensional data. Then we present our published large-scale dataset for this problem.

² <https://competition.aiops-challenge.com/home/competition/1484452272200032281>

³ <https://github.com/iopsai/iops/tree/master/evaluation>

⁴ <https://workshop.aiops.org>

Table 2. Comparison of KPI anomaly detection datasets

Dataset	Origin	Average Length	Average Anomaly Ratio
A	Real	211 969.75	0.022 7
WSD	Real	36 471.00	0.016
UCR	Real	55 935.25	0.003
TODS	Real	10 000.00	0.058
Yahoo	Real	1 561.22	0.007
NAB	Real	15 863.10	0.098

KPI: key performance indicator

NAB: Numanta Anomaly Benchmark

TODS: Time-series Outlier Detection System

WSD: Web Service Dataset

4.1 Background

To ensure close monitoring of system events, a wide range of structured logs are generated, which contain crucial information for diagnosing failures. These structured logs document events and encompass multiple fields that provide descriptive details. For instance, consider the example of an online shopping platform that generates structured logs to record orders. In the provided example displayed in Table 3, the Timestamp field indicates the occurrence time of the order, the Amount field specifies the monetary value associated with the order, the Region field denotes the customer's location, and the internet service provider (ISP) field identifies the network utilized by the customer.

Engineers aim to gather aggregated information from specific fields, such as the total dollar amount, to monitor the overall status of the system. These fields, which encompass aggregated values, are referred to as measures^[6,9]. On the other hand, the remaining fields that characterize events are referred to as attributes^[6-7,9,11].

By grouping event-level records within a monitoring interval by Region and ISP and aggregating Order Amount, engineers obtain multi-dimensional data^[6]. Table 4 presents one resulting snapshot, including actual and expected order amounts, which serves as direct input for root cause localization.

A failure in the system is indicated when the overall value of a measure deviates from the expected value, such as a decrease in the total dollar amount compared to the expected value. However, in practical scenarios, failures typically affect only specific attribute combinations, resulting in abnormal measure values^[7,47]. An attribute combination refers to a subset of the entire multi-dimensional data obtained by selecting records with specified attribute values, and its formal definition can be found in existing literature^[7,47]. In Table 4, for example, the measure values associated with {Region=Beijing} are substantially below their expected values. Such combinations indicate the scope of a failure and provide clues for diagnosis.

Table 3. Example of raw structured transaction logs

Order ID	Timestamp	Amount (CNY)	Region	ISP
T001	2021.11.11 10:00:01	102	Beijing	China Mobile
T002	2021.11.11 10:00:05	134	Beijing	China Unicom

ISP: Internet service provider

Table 4. Aggregated data grouped by Region and ISP for root-cause localization

Region	ISP	Amount (CNY)	Expected Amount
Beijing	China Mobile	32	64
Beijing	China Unicom	64	128
Shanghai	China Unicom	192	198
Guangdong	China Mobile	64	62
Zhejiang	China Unicom	13	13
Guangdong	China Unicom	1 283	1 347
Shanxi	China Unicom	128	141
Jiangsu	China Unicom	1 283	1 302
Tianjin	China Mobile	263	276

ISP: Internet service provider

4.2 Dataset

The majority of existing research studies on multi-dimensional data root cause localization^[7-16] utilize proprietary datasets for evaluation. Consequently, to advance the field, we have developed and made publicly available a dataset^[33] (referred to as Dataset B) based on real-world data obtained from Suning Ltd., a large online shopping platform. This dataset aims to facilitate research in this area.

Dataset B contains five attributes labeled as i, e, c, p, and l for confidentiality reasons. The measure is defined as the average number of orders per minute. The dataset comprises a total of 400 synthetic failures, each accompanied by ground-truth root cause labels, spanning eight weeks. Synthetic failures were used instead of real-world failures due to the difficulty in obtaining a substantial number of production failures with known ground-truth root causes.

The process of synthesizing failures involves six steps:

1) Anomaly injection: A time point is selected where the overall measure values are smooth.

2) Attribute selection: A random set of m attributes (where $m \in \{1, 2, 3, 4\}$) is chosen from the five available attributes.

3) Ground-truth root cause selection: n attribute combinations (where $n \in \{1, 2, 3, 4\}$) are randomly selected to represent the ground-truth root cause. These combinations correspond to the m selected attributes.

4) Modification of measure values: The measure values of the rows that satisfy the selected attribute combinations at the chosen time point are modified. The modified values are calculated using the Generalized Ripple Effect (GRE)

method^[47] with a random magnitude.

5) Addition of noise: Gaussian-distributed noise is added to the modified rows and other rows to simulate cases where GRE is not perfectly accurate and forecast residuals exist.

6) Synthetic failure validation: The synthetic failure is checked against the following conditions. If any of these conditions are triggered, the process returns to Step 2 to regenerate a failure at the same time point.

- Existence of other attribute combinations that share almost the same rows as the selected ground-truth attribute combinations.

- Injection of noise that is so significant that the overall measure value of unmodified records (from Step 3) becomes abnormal.

Among the 400 synthetic faults, 144 faults contain 1 root-cause attribute combination, 131 contain 2 combinations, 89 contain 3 combinations, and 36 contain 4 combinations. Furthermore, 96 faults involve 1 attribute, 136 involve 2 attributes, 108 involve 3 attributes, and 60 involve 4 attributes.

Utilizing Dataset B, we organized the second AIOps Challenge in 2019.⁵ The competition attracted 141 teams and 547 participants from both academia (39%) and industry (61%). The best-performing team achieved an F1-score of 0.959 3. Detailed solutions proposed by the top teams have been published in the workshop proceedings^[42].

Several existing studies^[6, 10] have utilized the data provided by the AIOps Challenge 2019. The B0-B4 datasets released with Squeeze^[6] are derived from the same original data as Dataset B and use a similar synthesis process.

As shown in Table 5, Dataset B contains 400 labeled synthetic failure cases generated from production measurements, whereas the RobustSpot evaluation dataset^[48] contains 135 manually investigated real-world anomaly cases. Both datasets have five attributes. While numerous studies in this field rely on private datasets that cannot be independently verified or reproduced, we introduce this dataset as a significant contribution to the research community. The severe shortage of public datasets has historically hindered reproducible research and fair comparative evaluations in root cause localization.

5 Failure Discovery and Diagnosis

In Sections 3 and 4, we presented our datasets for KPI anomaly detection and root cause localization for multi-dimensional data, respectively. These datasets serve as valuable resources for evaluating failure discovery and diagnosis algorithms, respectively. However, there are two important considerations. Firstly, engineers often require end-to-end automatic failure discovery and diagnosis. Secondly, while root cause attribute combinations provide important clues,

Table 5. Comparison of multi-dimensional data root cause localization datasets

Dataset	Origin	Case	Dimension
B	Real	400	5
RobustSpot	Real	135	5

they still necessitate further diagnosis to obtain detailed and fine-grained root causes.

To address these considerations, we have constructed a dataset for benchmarking both failure discovery and fine-grained diagnosis approaches. In this section, we provide an overview of the problem's background, followed by a detailed description of our dataset.

5.1 Background

A distributed system comprises multiple cooperating components, such as load balancers, web servers, application servers, and databases. To illustrate, we present the architecture of a production distributed system called sysA at China Mobile Zhejiang, a major Internet service provider (see Table 4). Each vertex in the figure represents a component, and these components can be categorized into five classes: Oracle Service Bus (OSB), service, database (DB), Docker, and operating system (OS).

To monitor the status of these components, engineers collect various KPIs related to their performance, such as queries per minute, success rate, and average response time. By applying KPI anomaly detection techniques to these metrics, we can identify failures within the distributed system and determine which components are exhibiting abnormal behavior. However, due to complex dependencies, a failure occurring in a root-cause component can propagate to other components, resulting in abnormal KPIs for those as well^[40].

To diagnose the failure, it is necessary to distinguish the root cause components from the other abnormal components. Additionally, to localize the fine-grained root causes of a failure, we need to identify the specific root cause metrics associated with the root cause components. Each component has multiple metrics available for monitoring, such as CPU utilization and network throughput. By identifying the root cause metrics, such as CPU usage, we can determine that the failure is caused by CPU exhaustion in that particular component.

In summary, our objective is to detect failures through KPI anomaly detection, and subsequently identify the root cause components and the specific metrics associated with the failure.

As introduced in Section 3, KPI anomaly detection has been extensively studied. The existing approaches to identifying root cause components^[31, 39 - 40, 49 - 53] can be classified into two categories: trace-based^[31, 39 - 40, 51 - 52] and dependency graph-based^[49 - 50, 53]. In addition, other notable ap-

⁵ <https://competition.aiops-challenge.com/home/competition/1484446614851493956>

proaches aim to identify root cause metrics. FluxRank^[54] formulates the metric ranking task as a learning-to-rank problem, where a machine learning model is trained to rank metrics by their importance for anomaly detection. MicroDiag^[55], CauseRank^[56] and CIRCA^[57] discover causal graphs among metrics and rank them by their causal impact on the observed anomalies.

5.2 Dataset

In the field of failure diagnosis, many studies rely on private datasets for evaluation, limiting the reproducibility and comparability of results. While some existing works have published datasets, such as those by Zhou et al.^[31] and Li et al.^[40], these datasets primarily consist of traces and do not support fine-grained failure diagnosis. Although there are public datasets available, such as Azure and Alibaba datasets that contain fine-grained metrics, they lack ground-truth information regarding failure time points and root causes. This hinders the thorough evaluation and benchmarking of failure discovery and fine-grained diagnosis approaches.

To address this gap and provide a comprehensive benchmark, we have constructed and published a dataset called Dataset C^[34]. This dataset has already been utilized by several existing works^[38–40, 58]. By leveraging Dataset C, researchers can evaluate their failure discovery and fine-grained diagnosis techniques in a more comprehensive and standardized way.

Dataset C is based on the architecture of sysA, as depicted in Fig. 4. This dataset contains three types of monitoring data: traces, KPIs, and metrics. In the dataset, user requests are initially sent to one of the OSB components, which then invoke other components such as service and DB. The entire execution process of a user request is called a trace. Each trace is recorded using spans, which represent units of work performed by each component of the distributed system, following the OpenTracing specification.

An example trace is presented in Table 6, where each row

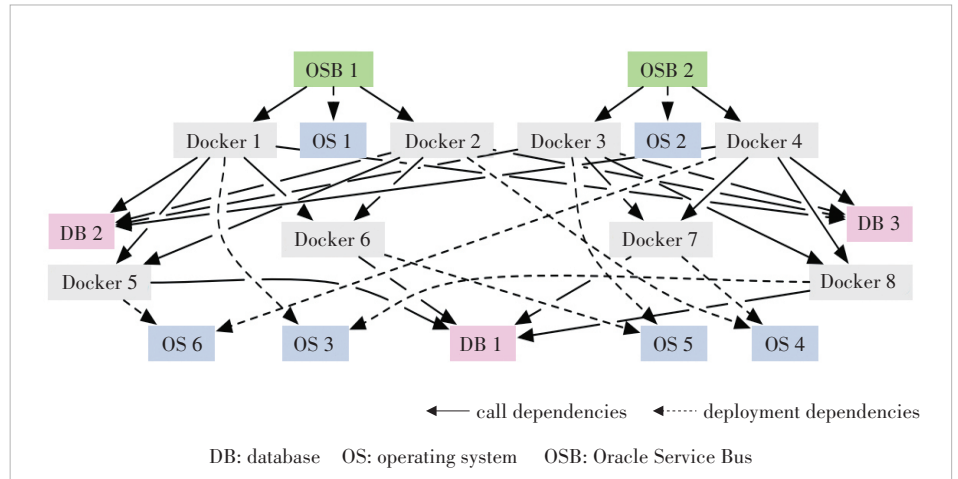


Figure 4. Components and their relationships, including call dependencies (solid) and deployment dependencies (dashed), in the distributed sysA

corresponds to a span. Each span has a unique identifier (id), a parent-span identifier (pid), and a trace identifier (traceId), which together describe the parent-child relationships within a trace. The remaining fields record the Unix start timestamp in milliseconds (startTime), duration in milliseconds (elapsedTime), response status (success), call type, and component identifier (cmdbId). In addition to traces and KPIs, Dataset C includes fine-grained metrics for each component.

By aggregating the attributes of spans within each component, KPIs can be derived to reflect the overall status of each component. The KPIs of a component, such as success rate and average response time, can be obtained by grouping spans based on the cmdbId field and aggregating relevant fields (e.g., success, elapsedTime). The system-level KPIs (i.e., KPIs of the OSB components) are already provided in the published dataset.

In addition to traces and KPIs, the dataset also includes fine-grained metrics for each component. These metrics can be categorized into four categories, as summarized in Table 7. These metrics provide detailed information about the performance and behavior of each component in the distributed system.

Using the traces, KPIs, and metrics provided in Dataset C, researchers can evaluate and develop techniques for failure discovery and fine-grained diagnosis realistically and comprehensively.

In Dataset C, we have included 169 injected failures with their corresponding ground-truth time points and root causes.

Table 6. An example trace in the dataset

callType	startTime(Unix)/ms	elapsedTime/ms	success	traceId	id	pid	cmdbId	serviceName
OSB	1590249600016	274	True	568385015	919979385015	None	os_021	osb_001
CSF	1590249600025	257	True	568385015	4eba31485915	919979385015	os_021	csf_001
RemoteProcess	1590249600026	254	True	568385015	82dad5063a15	4eba31485915	docke_003	csf_001

CSF: Client Service Framework

OSB: Oracle Service Bus

RemoteProcess: remote process call

These failures span one month. Injected failures are used instead of real-world failures due to the difficulty of obtaining a large number of real-world failures with accurately labeled root causes. In the sysA system, we have injected a total of 7 types of failures. These failure types and their descriptions are summarized in Table 8.

In 2020, the third AIOps Challenge⁶ was conducted based on Dataset C, which aimed at online failure detection and diagnosis. The competition attracted 141 teams and 517 participants from both academia and industry.

The challenge required participants to detect and diagnose the injected failures in Dataset C. For each failure, teams were allowed to submit at most two potential root-cause metrics. In cases of container network failures (loss and delay), as container-level network metrics were not collected, participants were asked to localize the root cause containers only.

To evaluate the submitted results, a precision score threshold of 0.5 was used. A submission for a failure was

Table 7. Metrics collected in Dataset C

Category	Metrics List
Docker	container_{thread_total, fgct, thread_idle, thread_used_pct, session_used, thread_running, fg, cpu_used, mem_used, fail_percent}
Linux	Agent_ping, Buffers_used, CPU_{idle_pct, user_time, system_time, iowait_time, util_pct}, Cache_used, Disk_{rd_ios, io_util, await, avgqu_sz, svctm, wr_kbs, rd_kbs, wr_ios}, FS_{used_space, used_pct, total_space, max_util, max_avail}, ICMP_ping, Incoming_network_traffic, Memory_{used_pct, used, available, free, available_pct, total}, Num_{of_processes, of_running_processes}, Outgoing_network_traffic, Page_{po, pi}, Processor_{load_5_min, load_1_min, load_15_min}, Received_{packets, errors_packets, queue}, Recv_total, Send_total, Sent_{errors_packets, queue, packets}, Shared_memory, Swap_used_pct, System_{wait_queue_length, block_queue_length}, Zombie_Process, ss_total
Oracle	ACS, AIOS, AWS, Asm_Free_Tb, CPU_{free_pct, Used_Pct}, Call_Per_Sec, DFParaWrite_Per_Sec, DbFile_Used_Pct, DbTime, Exec_Per_Sec, Hang, LFPParaWrite_Per_Sec, LFSync_Per_Sec, Logic_Read_Per_Sec, Login_Per_Sec, MEM_{real_util, Used, Total, Used_Pct}, New_{Tbs_Free_Gb, Tbs_Used_Pct}, On_Off_State, PGA_{Used_Pct, used_total}, Physical_Read_Per_Sec, Proc_{User_Used_Pct, Used_Pct}, Redo_Per_Sec, Row_Lock, SEQ_Used_Pct, SctRead_Per_Sec, SeqRead_Per_Sec, Sess_{Used_Undo, Used_Temp, Connect, Active}, Session_pct, TPS_Per_Sec, Tbs_{Used_Pct, Free_Gb}, TempTbs_Pct, Total_Tbs_Size, UndoTbs_Pct, Used_Tbs_Size, User_Commit, tnspring_result_time
Redis	Redis_key_count, blocked_clients, connected_clients, evicted_keys, expired_keys, instantaneous_{output_kbps, input_kbps, ops_per_sec}, keyspace_{hits, misses}, maxmemory, mem_fragmentation_ratio, redis_{ping, load}, rejected_connections, total_{connections_received, commands_processed}, used_{memory_peak, cpu_sys, memory_rss, cpu_user, memory}

considered valid only if met or exceeded this threshold. For each failure, valid submissions were ranked in ascending order of time to diagnose divided by the $F_{0.5}$ score (lower is better), and a result at rank i received $\max(10 - i + , 0)$ points. The best competition score was 755 over 169 scored failures. The public description does not specify how these 169 scored failures map to the 169 injected failures reported for Dataset C.

This challenge provided a platform for participants to demonstrate their skills in failure detection and diagnosis using Dataset C, further advancing research and development in AIOps.

As shown in Table 9, Dataset C differs from existing benchmarks in provenance and monitoring modalities. It provides aligned traces and metrics from a production system with injected-failure ground truth. DejaVu^[56] reports 601 failures across three production systems and one open-source benchmark, whereas Eadro^[58] reports 234 injection operations on two open-source testbeds. These counts are not directly comparable because the studies use different systems

Table 8. Failure injection in Dataset C

Component	Injection type	Description
Database	Close	Close the listening port of the target instance
Database	Session limit	Decrease the session limit of the target instance
Container	CPU stress (*)	Stress the CPU of the target container
Container	Network delay (\$)	Randomly delay packets on the target container
Container	Network loss (\$)	Randomly drop packets on the target container
Physical node	Network delay (\$)	Randomly delay packets on the target node
Physical node	Network loss (\$)	Randomly drop packets on the target node

Notes: stress-ng (*) and tc (\$) are injection tools.

Table 9. Comparison of failure discovery and diagnosis datasets

Dataset	Origin	Cases	Modality
Dataset C	Real	169	Trace/Metric
DejaVu	Testbed	601	Trace/Metric
Eadro	Testbed	117	Log/Metric/Trace
Nezha	Testbed	21	Log/Metric/Trace
SMAP, MSL	Real	105	Metric
TraceRCA	Testbed	33	Trace

MSL: Mars Science Laboratory

SMAP: Soil Moisture Active and Passive

⁶ <https://competition.aiops-challenge.com/home/competition/1484441527290765368>

and counting units.

6 Conclusions

AIOps has attracted significant attention from both academia and industry due to its potential. However, the lack of publicly available real-world and large-scale datasets remains a major obstacle to research progress. To address this gap, we have published three datasets covering KPI anomaly detection, multi-dimensional root cause localization, and failure discovery and diagnosis.

In addition, we have organized an annual algorithm competition based on these datasets, attracting numerous teams from around the world. This competition provides a platform for practitioners and researchers to advance AIOps techniques and enhance service reliability.

Moving forward, we will continue releasing datasets covering a wider range of AIOps scenarios and plan to persist in organizing the competition.

References

- [1] Moore S. Gartner: how to get started with AIOps [EB/OL]. [2024-04-15]. <https://www.gartner.com/smarterwithgartner/how-to-get-started-with-aiops>
- [2] Bu J, Liu Y, Zhang S, et al. Rapid deployment of anomaly detection models for large number of emerging KPI streams [C]//Proceedings of the 2018 IEEE International Performance, Computing and Communications Conference (IPCCC). IEEE, 2018: 1 – 8. DOI: 10.1109/PCCC.2018.8711315
- [3] Xu H, Chen W, Zhao N, et al. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications [C]//Proceedings of the 2018 World Wide Web Conference (WWW). ACM, 2018: 187 – 196. DOI: 10.1145/3178876.3185996
- [4] Liu D P, Zhao Y J, Xu H W, et al. Opprentice: towards practical and automatic anomaly detection through machine learning [C]//Proceedings of the 2015 Internet Measurement Conference. ACM, 2015: 211 – 224. DOI: 10.1145/2815675.2815679
- [5] Ma M, Zhang S, Chen J, et al. Jump-starting multivariate time series anomaly detection for online service systems [C]//Proceedings of the 2021 USENIX Annual Technical Conference. USENIX Association, 2021: 413 – 426. DOI: 10.5555/3472716.3472752
- [6] Li Z Y, Luo C Y, Zhao Y W, et al. Generic and robust localization of multi-dimensional root causes [C]//Proceedings of IEEE 30th International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2019: 47 – 57. DOI: 10.1109/ISSRE.2019.00015
- [7] Gu J, Luo C, Qin S, et al. Efficient incident identification from multi-dimensional issue reports via metaheuristic search [C]//Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020). ACM, 2020: 1234 – 1245. DOI: 10.1145/3368089.3409742
- [8] Ahmed F, Erman J, Ge Z H, et al. Detecting and localizing end-to-end performance degradation for cellular data services based on TCP loss ratio and round trip time [J]. IEEE/ACM transactions on networking, 2017, 25(6): 3709 – 3722. DOI: 10.1109/TNET.2017.2761758
- [9] Bhagwan R, Kumar R, Ramjee R, et al. Adtributor: revenue debugging in advertising systems [C]//Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14). USENIX Association, 2014: 43 – 55
- [10] Jing P K, Han Y N, Sun J Y, et al. AutoRoot: a novel fault localization schema of multi-dimensional root causes [C]//Proceedings of IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2021: 1 – 7. DOI: 10.1109/wcnc49053.2021.9417302
- [11] Lin Q, Lou J-G, Zhang H, et al. iDice: problem identification for emerging issues [C]//Proceedings of the 38th International Conference on Software Engineering (ICSE 2016). ACM, 2016: 214 – 224. DOI: 10.1145/2884781.2884824
- [12] Lin F, Muzumdar K, Laptev N P, et al. Fast dimensional analysis for root cause investigation in a large-scale service environment [J]. Proceedings of the ACM on measurement and analysis of computing systems, 2020, 4(2): 1 – 23. DOI: 10.1145/3392149
- [13] Qureshi M A, Qiu L L, Mahimkar A, et al. Multi-dimensional impact detection and diagnosis in cellular networks [C]//Proceedings of 16th International Conference on Mobility, Sensing and Networking (MSN). IEEE, 2020: 561 – 568. DOI: 10.1109/msn50589.2020.00093
- [14] Sun Y Q, Zhao Y J, Su Y, et al. HotSpot: anomaly localization for additive KPIs with multi-dimensional attributes [J]. IEEE access, 2018, 6: 10909 – 10923. DOI: 10.1109/ACCESS.2018.2804764
- [15] Wang H, Rong G P, Xu Y C, et al. ImpAPTr: a tool for identifying the clues to online service anomalies [C]//Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering. ACM, 2021: 1307 – 1311. DOI: 10.1145/3324884.3415301
- [16] Zhang X, Du C, Li Y F, et al. HALO: hierarchy-aware fault localization for cloud systems [C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. ACM, 2021: 3948 – 3958. DOI: 10.1145/3447548.3467190
- [17] Ma M H, Yin Z, Zhang S L, et al. Diagnosing root causes of intermittent slow queries in cloud databases [J]. Proceedings of the VLDB endowment, 2020, 13(8): 1176 – 1189. DOI: 10.14778/3389133.3389136
- [18] Zhao N, Chen J, Wang Z, et al. Real-time incident prediction for online service systems [C]//Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020). ACM, 2020: 315 – 326. DOI: 10.1145/3368089.3409745
- [19] Lin Q, Hsieh K, Dang Y, et al. Predicting node failure in cloud service systems [C]//Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2018). ACM, 2018: 858 – 868. DOI: 10.1145/3236024.3275511
- [20] Xu Y, Sui K, Yao R, et al. Improving service availability of cloud systems by predicting disk error [C]//Proceedings of the 2018 USENIX Annual Technical Conference (USENIX ATC 18). USENIX Association, 2018: 481 – 494. DOI: 10.5555/3277355.3277400
- [21] Chen Y J, Yang X, Lin Q W, et al. Outage prediction and diagnosis for cloud service systems [C]//Proceedings of World Wide Web Conference. ACM, 2019: 2659 – 2665. DOI: 10.1145/3308558.3313501
- [22] Shatnawi M, Hefeeda M. Real-time failure prediction in online services [C]//Proceedings of IEEE Conference on Computer Communications (INFOCOM). IEEE, 2015: 1391 – 1399. DOI: 10.1109/info-com.2015.7218516
- [23] Zhang K, Xu J, Min M R, et al. Automated IT system failure prediction: a deep learning approach [C]//Proceedings of the 2016 IEEE International Conference on Big Data (Big Data 2016). IEEE, 2016: 1291 – 1300. DOI: 10.1109/BigData.2016.7840760
- [24] Chen Z B, Kang Y, Li L Q, et al. Towards intelligent incident management: why we need it and how we make it [C]//Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ACM, 2020: 1487 – 1497. DOI: 10.1145/3368089.3417055
- [25] Chen J, He X, Lin Q, et al. An empirical investigation of incident triage for online service systems [C]//Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice. IEEE, 2019: 111 – 120. DOI: 10.1109/ICSE-SEIP.2019.00021
- [26] Chen J, He X, Lin Q, et al. Continuous incident triage for large-scale online service systems [C]//Proceedings of the 34th IEEE/ACM International

- tional Conference on Automated Software Engineering (ASE 2019). IEEE, 2019: 1126 – 1129. DOI: 10.1109/ASE.2019.00128
- [27] Lou J G, Lin Q W, Ding R, et al. Experience report on applying software analytics in incident management of online service [J]. Automated software engineering, 2017, 24(4): 905 – 941. DOI: 10.1007/s10515-017-0218-1
- [28] Delimitrou C, Kozyrakis C. Quasar: resource-efficient and QoS-aware cluster management [J]. ACM SIGPLAN notices, 2014, 49(4): 127 – 144. DOI: 10.1145/2644865.2541941
- [29] Cortez E, Bonde A, Muzio A, et al. Resource central: understanding and predicting workloads for improved resource management in large cloud platforms [C]//Proceedings of the 26th Symposium on Operating Systems Principles. ACM, 2017: 153 – 167. DOI: 10.1145/3132747.3132772
- [30] Laptev N, Amizadeh S, Flint I. Generic and scalable framework for automated time-series anomaly detection [C]//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1939 – 1947. DOI: 10.1145/2783258.2788611
- [31] Zhou X, Peng X, Xie T, et al. Latent error prediction and fault localization for microservice applications by learning from system trace logs [C]//Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ACM, 2019: 683 – 694. DOI: 10.1145/3338906.3338961
- [32] NetManAIOps. KPI-Anomaly-Detection: 2018 AIOps Challenge dataset [EB/OL]. [2024-04-15]. <https://github.com/netmanaiops/kpi-anomaly-detection>
- [33] NetManAIOps. MultiDimension-Localization: 2019 AIOps Challenge dataset [EB/OL]. [2024-04-15]. <https://github.com/NetManAIOps/MultiDimension-Localization>
- [34] NetManAIOps. AIOps-Challenge-2020-Data [EB/OL]. [2024-04-15]. <https://github.com/NetManAIOps/AIOps-Challenge-2020-Data>
- [35] Wang X R, Yin K L, Ouyang Q Y, et al. Identifying erroneous software changes through self-supervised contrastive learning on time series data [C]//Proceedings of IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2022: 366 – 377. DOI: 10.1109/issre55969.2022.00043
- [36] Ren H S, Xu B X, Wang Y J, et al. Time-series anomaly detection service at microsoft [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2019: 3009 – 3017. DOI: 10.1145/3292500.3330680
- [37] Zhao N W, Zhu J, Liu R, et al. Label-less: a semi-automatic labelling tool for KPI anomalies [C]//Proceedings of IEEE Conference on Computer Communications. IEEE, 2019: 1882 – 1890. DOI: 10.1109/info-com.2019.8737429
- [38] Cai Y, Han B, Li J, et al. ModelCoder: a fault model based automatic root cause localization framework for microservice systems [C]//Proceedings of IEEE/ACM 29th International Symposium on Quality of Service (IWQOS). IEEE, 2021: 1 – 6. DOI: 10.1109/iwqos52092.2021.9521318
- [39] Yu G B, Chen P F, Chen H Y, et al. MicroRank: end-to-end latency issue localization with extended spectrum analysis in microservice environments [C]//Proceedings of the Web Conference 2021. ACM, 2021: 3087 – 3098. DOI: 10.1145/3442381.3449905
- [40] Li Z Y, Chen J J, Jiao R, et al. Practical root cause localization for microservice systems via trace analysis [C]//Proceedings of IEEE/ACM 29th International Symposium on Quality of Service (IWQOS). IEEE, 2021: 1 – 10. DOI: 10.1109/iwqos52092.2021.9521340
- [41] Chen Z, Liu J, Su Y, et al. Adaptive performance anomaly detection for online service systems via pattern sketching [C]//Proceedings of the 44th International Conference on Software Engineering (ICSE 2022). ACM, 2022: 2009 – 2020. DOI: 10.1145/3510003.3510087
- [42] AIOps Workshop [EB/OL]. [2024-04-15]. <https://workshop.aiops.org>
- [43] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database [C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009). IEEE, 2009: 248 – 255. DOI: 10.1109/CVPR.2009.5206848
- [44] Notaro P, Cardoso J, Gerndt M. A survey of AIOps methods for failure management [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(6): 1 – 45. DOI: 10.1145/3483423
- [45] Zhang X, Kim J, Lin Q, et al. Cross-dataset time series anomaly detection for cloud systems [C]//Proceedings of the 2019 USENIX Annual Technical Conference (USENIX ATC 19). USENIX Association, 2019: 1063 – 1076. DOI: 10.5555/3358807.3358903
- [46] Lavin A, Ahmad S. Evaluating real-time anomaly detection algorithms: the numenta anomaly benchmark [C]//Proceedings of IEEE 14th International Conference on Machine Learning and Applications (ICMLA). IEEE, 2015: 38 – 44. DOI: 10.1109/ICMLA.2015.141
- [47] Li Z, Chen J, Chen Y, et al. Generic and robust root cause localization for multi-dimensional data in online service systems [J]. Journal of systems and software, 2023, 203: 111748. DOI: 10.1016/j.jss.2023.111748
- [48] Li Z, Zhao N, Li M, et al. Actionable and interpretable fault localization for recurring failures in online service systems [C]//Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2022). ACM, 2022: 996 – 1008. DOI: 10.1145/3540250.3549162
- [49] Kim M, Sumbaly R, Shah S. Root cause detection in a service-oriented architecture [J]. ACM SIGMETRICS performance evaluation review, 2013, 41(1): 93 – 104. DOI: 10.1145/2494232.2465753
- [50] Ma M, Xu J, Wang Y, et al. AutoMAP: diagnose your microservice-based web applications automatically [C]//Proceedings of the 2020 World Wide Web Conference (WWW 2020). ACM, 2020: 320 – 330. DOI: 10.1145/3366423.3380115
- [51] Lin J, Chen P, Zheng Z. Microscope: pinpoint performance issues with causal graphs in micro-service environments [C]//Proceedings of the 16th International Conference on Service-Oriented Computing (ICSOC 2018). Springer, 2018: 3 – 20. DOI: 10.1007/978-3-030-03596-9_1
- [52] Gan Y, Liang M, Dev S, et al. Sage: practical and scalable ML-driven performance debugging in microservices [C]//Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2021). ACM, 2021: 135 – 151. DOI: 10.1145/3445814.3446700
- [53] Yu Q, Pei C, Hao B, et al. CMDiagnositor: an ambiguity-aware root cause localization approach based on call metric data [C]//Proceedings of the ACM Web Conference 2023 (WWW 2023). ACM, 2023: 2937 – 2947. DOI: 10.1145/3543507.3583347
- [54] Liu P, Chen Y, Nie X, et al. FluxRank: a widely-deployable framework to automatically localize root cause machines for software service failure mitigation [C]//Proceedings of the 2019 IEEE 30th International Symposium on Software Reliability Engineering. IEEE, 2019: 35 – 46. DOI: 10.1109/ISSRE.2019.00014
- [55] Wu L, Tordsson J, Bogatinovski J, et al. MicroDiag: fine-grained performance diagnosis for microservice systems [C]//Proceedings of IEEE/ACM International Workshop on Cloud Intelligence (CloudIntelligence). IEEE, 2021: 31 – 36. DOI: 10.1109/cloudintelligence52565.2021.00015
- [56] Lu X, Xie Z, Li Z, et al. Generic and robust performance diagnosis via causal inference for OLTP database systems [C]//Proceedings of the 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid 2022). IEEE, 2022: 655 – 664. DOI: 10.1109/CCGrid54584.2022.00076
- [57] Li M, Li Z, Yin K, et al. Causal inference-based root cause analysis for online service systems with intervention recognition [C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022). ACM, 2022: 3230 – 3240. DOI: 10.1145/3534678.3539041
- [58] Lee C, Yang T, Chen Z, et al. Eadro: an end-to-end troubleshooting

framework for microservices on multi-source data [C]//Proceedings of the 45th IEEE/ACM International Conference on Software Engineering (ICSE 2023). IEEE, 2023: 1750 – 1762. DOI: 10.1109/ICSE48619.2023.00150

Biographies

Li Zeyan received his PhD degree in computer science and technology from Tsinghua University, China in 2023, where he was a member of the Tsinghua NetMan Lab under the supervision of Prof. Pei Dan. He received his bachelor's degree in computer science and technology from Tsinghua University in 2018. His research interests focus on Artificial Intelligence for IT Operations (AIOps), particularly intelligent fault localization, root cause analysis, and reliability management for large-scale online service systems.

Han Jing (han.jing28@zte.com.cn) is a researcher and engineer at ZTE Corporation, where she develops next-generation intelligent network architectures, 6G wireless intelligence, and automated software maintenance systems.

Zhao Nengwen received her PhD degree in computer science and technology from Tsinghua University, China in 2022 under the supervision of Prof. Pei Dan, and her bachelor's degree in computer science and technology from Wuhan University, China in 2017. Her research interests include Artificial Intelligence for IT Operations (AIOps), software maintenance, and reliability management, with particular emphasis on anomaly detection, alert management, log anomaly detection, and software change analysis in large-scale online service systems.

Zhang Shenglin is a Professor at the College of Software, Nankai University, China and an affiliated faculty member of the Tsinghua NetMan Lab. He received his PhD degree in computer science and technology from Tsinghua University, China under the supervision of Prof. Pei Dan and Prof. Liu Ying, and his bachelor's degree in network engineering from Xidian University, China in 2012. During his doctoral studies, he was a visiting scholar at the School of Computer Science, Georgia Institute of Technology, USA. His research focuses on Artificial Intelligence for IT Operations (AIOps), including failure detection, diagnosis and prediction, root cause analysis, log and multi-modal observability analysis, and reliability management for large-scale service and cloud systems.

Sun Yongqian is an Associate Professor at Nankai University, China and an affiliated faculty member of the Tsinghua NetMan Lab. He received his PhD

degree from the Department of Computer Science and Technology at Tsinghua University, China under the supervision of Prof. Pei Dan and Prof. Zhao Youjian, and his bachelor's degree in statistics from Northwestern Polytechnical University, China in 2012. His research interests center on Artificial Intelligence for IT Operations (AIOps), including anomaly detection, failure diagnosis, root cause analysis, failure prediction, and intelligent operations for microservice and large-scale software systems.

Chen Pengfei is a Professor and doctoral supervisor at the School of Computer Science and Engineering, Sun Yat-sen University, China. He received his PhD degree in computer science from Xi'an Jiaotong University, China in 2016. His research interests include Artificial Intelligence for IT Operations (AIOps), distributed systems, cloud computing, microservices, software reliability, and computer networks, with particular focus on automated performance diagnosis, root cause analysis, anomaly detection, and intelligent management of cloud-native and microservice systems.

Wen Xidao was a postdoctoral researcher at Tsinghua University, China and a member of the Tsinghua NetMan Lab, completing his postdoctoral work in 2022. His research interests include Artificial Intelligence for IT Operations (AIOps), data mining, machine learning, time-series analysis, anomaly detection, fault diagnosis, and intelligent operations for large-scale service systems. He has collaborated extensively on research in multivariate time-series anomaly detection, software change analysis, root cause diagnosis, AIOps benchmarks, and large language models for IT operations.

Ma Minghua is a Senior Researcher at Microsoft Research in Redmond, USA. He received his PhD degree in computer science and technology from Tsinghua University, China in 2021 under the supervision of Prof. Pei Dan. From 2019 to 2020, he was a visiting scholar at the School of Computer Science, Georgia Institute of Technology, USA. His research interests lie at the intersection of software engineering, systems, and artificial intelligence, with a particular focus on Cloud Intelligence and Artificial Intelligence for IT Operations (AIOps).

Pei Dan is a Professor in the Department of Computer Science and Technology, Tsinghua University, China and the leader of the Tsinghua NetMan Lab. His current research focuses on Artificial Intelligence for IT Operations (AIOps) for critical infrastructure and embodied systems, as well as time-series intelligence. Before joining Tsinghua University, he was a researcher at AT&T Labs – Research in the United States. He is a Senior Member of both ACM and IEEE and has extensive research experience in networked systems, large-scale service reliability, intelligent operations, root cause analysis, anomaly detection, and time-series intelligence.

Reconfigurable Metasurface for Near-Field Efficient Wireless Power Transfer System



Wang Yunfeng¹, Xu Zhiyu², Zhang Hongtao¹

(1. Key Lab of Universal Wireless Communications, Ministry of Education of China, Beijing University of Posts and Telecommunications, Beijing 100876, China;

2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202603009

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260901.1008.002.html>, published online September 1, 2026

Manuscript received: 2024-12-13

Abstract: Wireless power transfer (WPT) offers a promising solution for powering distributed devices in smart cities, significantly simplifying system complexity and accelerating urban automation. In traditional WPT systems, far-field schemes are costly, energy-intensive, and inefficient, while near-field transmission offers higher efficiency but suffers from limited distance. To address the limited transmission distance of near-field WPT, this paper derives an electromagnetic model for WPT using reflective metasurfaces and presents the design of a high-performance 3-bit reconfigurable intelligent surface (RIS). Based on theoretical models, we investigate the effects of phase quantization and array size on focusing efficiency. Simulation results guide the subsequent fabrication of the RIS prototype. Experimental measurements on the fabricated prototype validate the beam-focusing effect and evaluate the overall WPT efficiency.

Keywords: reconfigurable intelligent surface; near-field beam focusing; wireless power transfer

Citation (Format 1): Wang Y F, Xu Z Y, Zhang H T. Reconfigurable metasurface for near-field efficient wireless power transfer system [J]. *ZTE Communications*, 2026, 24(3): 70 – 81. DOI: 10.12142/ZTECOM.202603009

Citation (Format 2): Y. F. Wang, Z. Y. Xu, and H. T. Zhang, “Reconfigurable metasurface for near-field efficient wireless power transfer system,” *ZTE Communications*, vol. 24, no. 3, pp. 70 – 81, Sept. 2026. doi: 10.12142/ZTECOM.202603009.

1 Introduction

Wireless power transfer (WPT) enables the transfer of electrical energy from the power source to a load through electromagnetic waves, without physical wire connections. It offers features such as on-demand charging, non-contact operation, safety, reliability, and simultaneous charging of multiple devices, eliminating cables and greatly enhancing convenience and flexibility.

WPT technology has the potential to revolutionize traditional charging methods, significantly improving the flexibility, convenience, and reliability of equipment such as electric vehicles and smart power systems. It can drive innovation across sectors including electric vehicles, smart grids, medical devices, home appliances, and aerospace. At the same time, it ensures secure and stable grid operation and enhances energy efficiency, offering vast market potential.

Unlike electromagnetic induction charging, which only supports short-range power transfer, WPT based on radio frequency (RF) energy harvesting can operate over longer distances, providing greater flexibility for power-driven devices^[1-2]. A complete WPT system requires determining the

receiver's location and adjusting the transmission system accordingly, followed by signal emission, transmission, and reception. Current research on transmission systems mainly focuses on three categories: far-field RF transmitters^[3-4], fixed antennas^[5-6], and energy transfer assisted by metasurfaces^[7-9]. However, far-field RF power transmission suffers from low efficiency and limited application scenarios. It is heavily affected by path loss, and the electric field distribution is diffuse, resulting in low efficiency. It can only power low-energy devices through specialized collection and processing at the receiver. In contrast, near-field radiation offers higher transmission efficiency, better meeting the functional needs of indoor devices^[10-12]. The focusing efficiency of metasurfaces depends primarily on phase control accuracy, the reflection loss coefficient, and phase compensation precision. Phase quantization accuracy is limited by the unit resonator component design. Current adjustable metasurface prototypes commonly use PIN diodes^[13-15] and varactor diodes^[16] for phase tuning, but PIN-diode-based designs are typically limited to discrete phase states, while varactor-diode-based designs may face a tradeoff between the achievable phase-tuning range and tuning loss, requiring further improvement in transmission efficiency.

Over the past two decades, research on metamaterials and

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20240000007.

metasurfaces has experienced explosive growth^[17]. Metasurfaces are artificial materials designed to achieve electromagnetic phenomena that are either unavailable or difficult to obtain in nature. With their wave manipulation characteristics, they are implemented as an array of specially arranged scatterers in a two-dimensional periodic layout^[17–18]. Subsequently, the theory and applications of metasurfaces have been developed in tandem within and beyond the antenna and propagation communities^[19–20]. Groundbreaking work in the field of gradient metasurfaces^[21], along with the manipulation of amplitude and phase of incident wavefronts using nanostructures^[22], has sparked significant interest in the control of light wavefronts with metasurface devices. The engineering concepts of phased arrays, reflective beams, and emission arrays have further highlighted the scientific and commercial value of these novel devices, increasing interest in metamaterials and metasurfaces^[23–25].

The transmission of microwave energy through antenna arrays^[26] offers high efficiency, but the supporting equipment is complex and costly due to the use of traditional antennas. Energy focusing using a fixed metasurface^[10] shows good performance at a preset focusing position, but the fixed unit design makes it difficult to achieve consistent and efficient energy transfer when the focusing position changes. By using PIN diodes, 1-bit and 2-bit reflective WPT reconfigurable intelligent surface (RIS) can be achieved^[11, 27]. However, PIN diodes introduce design complexity and limit control accuracy, which reduces energy transfer efficiency. This paper proposes a 3-bit RIS unit based on low-loss varactor diodes, which significantly reduces the antenna cost and control complexity of the approach in Ref. [26] and improves WPT efficiency compared to lower-bit phase quantization in Refs. [11] and [27]. Compared to Ref. [10], it can achieve consistent energy transfer efficiency at different focusing positions. Our research on quantization accuracy and array size has guided the design of an efficient RIS for energy transmission.

In this paper, we propose a feasible and efficient WPT strategy that utilizes reflective RIS to achieve near-field energy-focused transmission. The rest of this paper is organized as follows. We discuss the efficient near-field WPT strategy using RIS and establish a transmission efficiency model in Section 2. Section 3 presents the design of the 3-bit RIS and simulations of transmission efficiency based on it. We also fabricate prototypes and present actual measurement results in Section 4. Finally, Section 5 concludes the paper.

2 RIS-Aided Near-Field WPT Framework

In WPT systems, a key metric is energy transmission efficiency, which is significantly influenced by the power transmission beam. This paper proposes an efficient power beam-forming strategy that integrates the location information of the device to be charged and the RIS. The strategy automatically calculates and allocates high-gain near-field energy beams to

meet the energy transmission requirements across different scenarios. Fig. 1 illustrates the application scenario of an RIS in an indoor WPT system.

2.1 Reflective RIS and Numerical Model

In recent years, there has been growing interest in developing dynamic metasurfaces for constructing advanced reconfigurable devices. The concept of RIS has been introduced, enabling real-time control over electromagnetic wave propagation. Similar to digital circuit technology, the electromagnetic responses of metasurface elements are represented by discrete digital codes. By modifying these code sequences, functional control of both far-field and near-field electromagnetic behavior can be achieved.

A metasurface can be defined as an array of polarizable elements that induce discontinuities in electric and magnetic fields on both sides of the surface. The macroscopic description of metasurfaces aims to express these field discontinuities as functions of the surface-averaged electric and magnetic polarization densities. This macroscopic approach replaces the detailed spatial distribution of individual elements with analytical relations between the electric and magnetic fields on either side of the surface.

The homogenization of the surface treats it as having continuous electric and magnetic polarization densities. This homogenization incorporates the unit cell's size, shape, and material properties directly into these analytical relations. This provides essential design guidance for metasurface elements.

The analytical relations linking the fields on both sides of the surface to the homogeneous polarization densities are known as the Generalized Sheet Transition Conditions (GSTCs)^[28]. This macroscopic description allows the polarization densities to be represented using magnetizability tensors instead of polarizability, relying only on the averaged fields. These magnetizability tensors, much like dielectric permittivity and magnetic permeability for volumetric metamaterials,

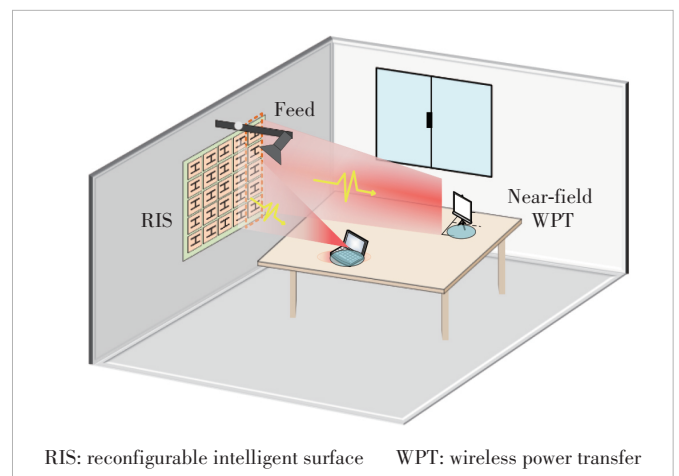


Figure 1. Application scenarios of RIS in a WPT system

act as effective parameters for the metasurface. They depend on the unit cell's topology, geometry, spacing, material properties, and wavelength. The electric ($\mathbf{P}$) and magnetic ($\mathbf{M}$) surface polarization densities can be formulated as follows:

$$\mathbf{M}(x, y) = \sqrt{\varepsilon/\mu} \bar{\chi}_{me}(x, y) \mathbf{E}_{av}(x, y) + \bar{\chi}_{mm}(x, y) \mathbf{H}_{av}(x, y) \quad (1),$$

$$\mathbf{P}(x, y) = \varepsilon \bar{\chi}_{ee}(x, y) \mathbf{E}_{av}(x, y) + \sqrt{\mu\varepsilon} \bar{\chi}_{em}(x, y) \mathbf{H}_{av}(x, y) \quad (2),$$

where ε and μ are vacuum permittivity and permeability, respectively; $\bar{\chi}_{ab}(x, y)$, for $a = \{e, m\}$ and $b = \{e, m\}$, are the electric-electric, electric-magnetic, magnetic-electric, and magnetic-magnetic surface-averaged (or simply surface) susceptibility dyadics, which are the macroscopic quantities of interest for a simple but accurate synthesis and analysis of metasurfaces; $\mathbf{E}_{av}$ and $\mathbf{H}_{av}$ are the surface-averaged electric and magnetic fields, where the average is computed over a small surface area of the order of the wavelength (including one or more unit cells depending on the setup).

Once the average reflection coefficient of the metasurface is determined, we incorporate analytical theories of incident, reflected, and surface electromagnetic waves to meet various performance requirements. This approach facilitates the derivation of a general susceptibility function, which helps optimize the design parameters of the unit cells^[29]. These analytical frameworks enable fine-tuning of the metasurface characteristics to achieve desired functionalities, ensuring that it performs effectively in different scenarios. The surface electromagnetic field results from the electromagnetic discontinuity introduced by the metasurface structure. Its analytical expression can be derived from the Generalized GSTCs. These conditions relate the fields on either side of the metasurface to the surface polarization densities, thereby providing a framework for understanding how the metasurface interacts with incident electromagnetic waves and influences the fields in its vicinity.

$$\begin{aligned} E_{x,y}^{rt}(x, y) &\propto \mathbf{s}_{E,x,y}^{rt}(x, y) E_x^i(x, y) + \mathbf{s}_{E,x,y}^{rt}(x, y) E_y^i(x, y) + \\ &\mathbf{s}_{H,x,y}^{rt}(x, y) H_x^i(x, y) + \mathbf{s}_{H,x,y}^{rt}(x, y) H_y^i(x, y) \end{aligned} \quad (3),$$

$$\begin{aligned} H_{x,y}^{rt}(x, y) &\propto \xi_{E,x,y}^{rt}(x, y) E_x^i(x, y) + \xi_{E,x,y}^{rt}(x, y) E_y^i(x, y) + \\ &\xi_{H,x,y}^{rt}(x, y) H_x^i(x, y) + \xi_{H,x,y}^{rt}(x, y) H_y^i(x, y) \end{aligned} \quad (4),$$

where the superscripts i, r and t denote the incident, reflected, and transmitted components of the fields, respectively; $E_{x,y}(x, y)$ and $H_{x,y}(x, y)$ represent the tangential components of the electric and magnetic fields, respectively; $\mathbf{s}_{\{E,H\},\{x,y\}}^{rt}(x, y)$ and $\xi_{\{E,H\},\{x,y\}}^{rt}(x, y)$, derived from a matrix that depends solely on the metasurface structure (specifically the surface magnetic susceptibility function), express each tangential component of

the surface electromagnetic fields on both sides of the metasurface as a linear combination of the tangential components of the incident field. In the case of perfect anomalous reflection, analyzing the reflective surface electromagnetic field provides valuable insights for the design of broadband metasurface structures, enabling precise control over the reflected waves and enhancing metasurface performance.

Under the assumption of perfect anomalous reflection, conditions $E_x^t(x, y) = E_y^t(x, y) = 0$ and $H_x^t(x, y) = H_y^t(x, y) = 0$ exist. The calculations for $E_y^r(x, y)$ and $H_x^r(x, y)$ are based on the conventional scenario. Here, $\chi_{ee}^{yy}(x, y)$ and $\chi_{mm}^{xx}(x, y)$ represent the surface-averaged electric and magnetic susceptibility functions. Due to $\chi_{ee}^{yy}(x, y) \chi_{mm}^{xx}(x, y) = -4/k^2$, the corresponding two constitutive relations in the reflected electromagnetic field are not linearly independent. This implies that multiple electromagnetic field transformations satisfy the constitutive relations. Considering a solution of the form $H_x^r(x, y) = e(x, y) E_y^r(x, y)$, where $e(x, y)$ is a given function such that $E_y^r(x, y)$ and $H_x^r(x, y)$ satisfy Maxwell's equations, we obtain an explicit expression for the reflected electromagnetic field as a function of the incident electromagnetic field under these assumptions:

$$\begin{aligned} E_y^r(x, y) &= \frac{-\frac{j\omega\varepsilon}{2} \chi_{ee}^{yy}(x, y)}{e(x, y) + \frac{j\omega\varepsilon}{2} \chi_{ee}^{yy}(x, y)} E_y^i(x, y) - \\ &\frac{1}{e(x, y) + \frac{j\omega\varepsilon}{2} \chi_{ee}^{yy}(x, y)} H_x^i(x, y) \end{aligned} \quad (5),$$

$$\begin{aligned} H_x^r(x, y) &= \frac{-\frac{j\omega\varepsilon}{2} e(x, y) \chi_{ee}^{yy}(x, y)}{e(x, y) + \frac{j\omega\varepsilon}{2} \chi_{ee}^{yy}(x, y)} E_y^i(x, y) - \\ &\frac{e(x, y)}{e(x, y) + \frac{j\omega\varepsilon}{2} \chi_{ee}^{yy}(x, y)} H_x^i(x, y) \end{aligned} \quad (6).$$

From the above expressions, the reflected electric and magnetic fields are represented as linear combinations of the incident electric and magnetic fields. Generally, there is cross-coupling between the electromagnetic fields. For any given matrix, its pseudoinverse always exists and is unique. This means that for a given metasurface structure (represented by the surface magnetic susceptibility function and the incident field), one can always find the corresponding reflected and transmitted fields. Therefore, the metasurface structure can be reverse-engineered from the reflected field, providing guidance for unit cell design.

2.2 Near-Field WPT Strategy Using RIS

To achieve near-field energy focusing using a reflective metasurface, a numerical model is first established to compute the electric field E in the Fresnel region, assuming the metasurface consists of $I \times J$ reflective units, each of size $px \times py$. The emitting source is located at point (x_f, y_f, z_f) , and the target focal point is at point (x_d, y_d, z_d) . Fig. 2 shows a schematic diagram of phase compensation for reflective focusing metasurfaces.

Assuming polarization along the y -axis in the model, the scattering E -field in the Fresnel region can be represented using the plane wave angular spectrum (PWAS) method^[11]:

$$E_y(x, y, z) = \frac{j}{2} \exp(-jk_0 z) \cdot \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} \left\{ E_t(x_i, y_j) \cdot \left[C(t_2) - jS(t_2) \right] - \left[C(t_1) - jS(t_1) \right] \right\} \cdot \left\{ \left[C(t_2') - jS(t_2') \right] - \left[C(t_1') - jS(t_1') \right] \right\} \quad (7)$$

where $k_0 = 2\pi/\lambda$ is the wave number in free space, $E_t(x_i, y_j)$ is the tangential E -field at the center of each unit, and $C(u)$ and $S(u)$ are the Fresnel integrals, which can be written as:

$$C(u) = \int_0^u \cos\left(\frac{\pi}{2} t^2\right) dt, \quad S(u) = \int_0^u \sin\left(\frac{\pi}{2} t^2\right) dt$$

$$t_1 = \sqrt{\frac{k_0}{\pi z}} \left(\frac{p_x}{2} + x - mp_x + \frac{(M-1)p_x}{2} \right)$$

$$t_2 = \sqrt{\frac{k_0}{\pi z}} \left(-\frac{p_x}{2} + x - mp_x + \frac{(M-1)p_x}{2} \right)$$

$$t_1' = \sqrt{\frac{k_0}{\pi z}} \left(\frac{p_y}{2} + y - np_y + \frac{(N-1)p_y}{2} \right)$$

$$t_2' = \sqrt{\frac{k_0}{\pi z}} \left(-\frac{p_y}{2} + y - np_y + \frac{(N-1)p_y}{2} \right) \quad (8)$$

The representation of each unit's $E_t(x_i, y_j)$ is given as follows:

$$E_t(x_i, y_j) = A_{ij} \exp(j\varphi_{ij}) \cdot E_0(x_i, y_j) \times \exp\left(jk_0 \sqrt{(x_f - x_i)^2 + (y_f - y_j)^2 + z_f^2}\right) \quad (9)$$

where A_{ij} and φ_{ij} represent the reflection amplitude and phase shift of the ij -th unit cell, respectively; $E_0(x_i, y_j)$ denotes the amplitude of the incident tangential electric field from the spatial source at the location of the n -th unit cell. To generate a focused spot at the desired position, the phase shifts of all units must align coherently at the target location.

According to electromagnetic wave propagation theory and

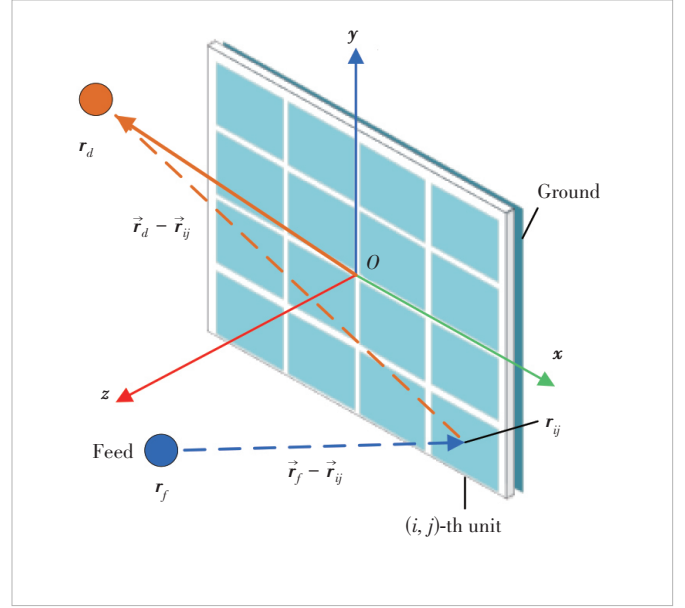


Figure 2. Schematic diagram of phase compensation for reflective focusing metasurfaces

the principle of coherent beam superposition, the phase equals the product of the wavenumber and the optical path. The optical path difference between each unit on the reflective focusing surface and the target focal point can be expressed as $|\vec{r}_d - \vec{r}_{ij}|$, as shown in Fig. 2. Assuming the reflective focusing surface serves as the reference phase plane, the phase distribution required to generate the focus is located behind the reference plane wave. Therefore, the phase distribution necessary for focusing in space can be expressed as:

$$\varphi_d(x_i, y_j) = \exp\left(-jk_0 |\vec{r}_d - \vec{r}_{ij}|\right) \quad (10)$$

Using conventional horn-fed excitation, the amplitude distribution on the reflective focusing surface is predetermined, as is the phase excitation distribution. When designing the reflective focusing surface, it is essential to account for the initial phase and amplitude distributions. Assuming the reflective focusing surface serves as the phase reference plane, the phase distribution of the incident feed on the metasurface can be expressed as:

$$\varphi_f(x_i, y_j) = \exp\left(jk_0 |\vec{r}_f - \vec{r}_{ij}|\right) \quad (11)$$

The required phase shift distribution for the reflective focusing surface is the difference between the focusing phase distribution necessary to direct the energy towards the focal point and the initial phase distribution of the incident feed on the metasurface. Thus, it can be expressed as:

$$\Delta\varphi_{ij}(x_i, y_j) = \varphi_d(x_i, y_j) - \varphi_f(x_i, y_j) \quad (12)$$

By setting the feed source position, the focal point, and the arrangement of all units on the metasurface, the phase compensation can be calculated through optical path differences and assigned to the metasurface elements. This enables energy synthesis at the desired focal point.

Using the above formulas and incorporating position information, a phase compensation matrix can be computed. The electric field amplitude and phase are then superimposed on the focusing surface to simulate the energy focusing effect of the reflective metasurface. Additionally, the impact of quantization bit depth and array aperture on focusing performance is preliminarily explored, providing insights for subsequent element and array design.

2.3 Quantization and Aperture Analysis

Focusing efficiency is a crucial metric that directly reflects the effectiveness of focusing-based WPT systems. In this section, the efficiency η is defined as the ratio of the total energy P_d within the focused area at the reference plane to the incident power P_f from the feed source on the metasurface:

$$\eta = \frac{P_d}{P_f} \quad (13).$$

The focal area is defined as the circular region on the focal plane that encompasses the point of maximum electric field intensity and the surrounding area with an intensity drop of 3 dB, as illustrated in Fig. 3. The total energy within the focal area P_d can be obtained by integrating the Poynting vector over this region. Similarly, the incident power is calculated by integrating the Poynting vector over the plane of the receiving horn antenna.

Existing research indicates that in near-field beam steering

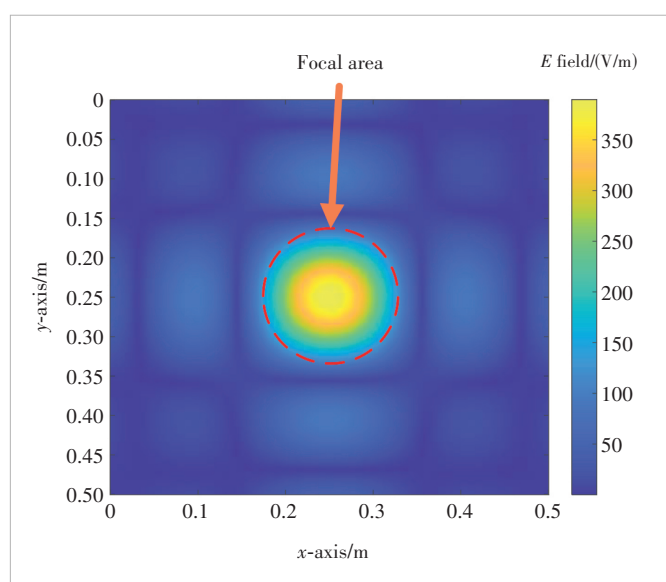


Figure 3. Electric field distribution of the metasurface focal area

scenarios, a 1-bit quantization scheme results in a gain reduction of approximately 3 dB^[11]. Increasing the quantization precision with more bits can mitigate this gain loss. Thus, a higher phase quantization resolution improves the focusing efficiency of the reflective focusing metasurface. However, the control overhead introduced by multi-bit quantization is significant and will substantially increase system costs. Therefore, it is crucial to balance the number of quantization bits against the focusing efficiency. This section illustrates this issue via simulation.

We set the unit size to 25 mm × 25 mm with an incident wave frequency of 5.8 GHz. A 20 × 20 array forms a 0.5 m × 0.5 m aperture, with the focal point fixed at the center, 1 m away from the metasurface. By calculating the compensation phases for each unit, a non-quantized compensation phase matrix is obtained. The matrix is then quantized to 1-bit, 2-bit, 3-bit, and 4-bit precision. The strongest electric field value at the focal point is compared for each quantization level (Fig. 4).

From Fig. 4, as the quantization precision increases from 1 bit to 4 bit, the electric field intensity at the focal center gradually rises (all normalized to the maximum under non-quantized control), but the rate of increase diminishes. This indicates a trend towards saturation. Specifically, increasing from 1 bit to 2 bit improves the focal intensity by 47.6%; from 2 bit to 3 bit, by 7.4%; and the 4-bit improvement is smaller. However, increasing the number of control bits significantly increases control link bandwidth requirements and system costs. After weighing performance against cost, this study adopts a 3-bit quantization scheme.

To determine the influence of array size on focusing performance, we simulate arrays of different sizes. Fig. 5 shows the simulated focusing effect. In the simulations, the array sizes are set to 10 × 10 units, 20 × 20 units, and 40 × 40 units, with dimensions of 0.25 m × 0.25 m, 0.5 m × 0.5 m, and 1 m × 1 m, respectively, focusing at distances of 1 m and 5 m from the metasurface center. Since the theoretical simulations do not account for path loss, the focused power can only be compared at the same distance. Non-quantized phase is used as the compensation phase, and the influence of the feeding path is ignored, assuming the reflected power of each unit is uniform across all array sizes.

From Fig. 5 and the focal area definition, we determine the focal spot sizes. At a focusing distance of 1 m, the focal spot sizes for the 10 × 10, 20 × 20, and 40 × 40 arrays are 19 cm × 19 cm, 9.5 cm × 9.5 cm, and 5.2 cm × 5.2 cm, respectively. At a focusing distance of 5 m, the focal spot size of the 10 × 10 array exceeds the size of the transmitting array, while those for the 20 × 20 and 40 × 40 arrays are 48 cm × 48 cm and 23.2 cm × 23.2 cm, respectively.

In practical WPT applications, the energy receiving area is typically fixed. Reducing the focal spot size increases the received power. However, increasing the array aperture multi-

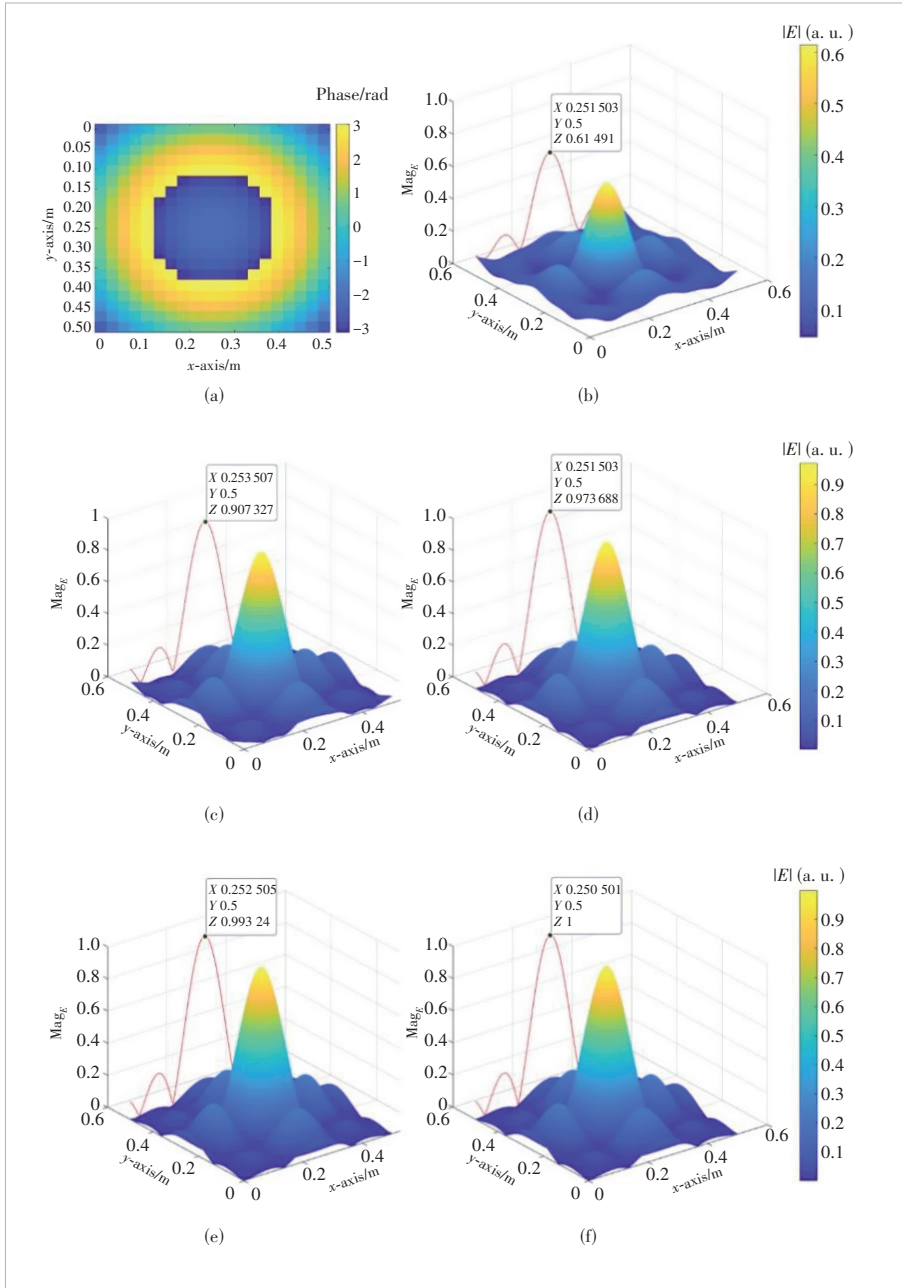


Figure 4. Phase compensation and normalized electric-field distributions: (a) non-quantized compensation phase distribution; (b) 1-bit quantization; (c) 2-bit quantization; (d) 3-bit quantization; (e) 4-bit quantization; (f) non-quantized electric-field distribution (reference)

plies the number of metasurface units, leading to a significant rise in costs. Thus, it is essential to balance the final array size against the transmission distance, aperture size, and focusing efficiency.

From Fig. 5, doubling the array size significantly reduces the focal size. At closer distances, the large-aperture gain of larger arrays is evident. For example, at 1 m, the 40×40 array has 300% more elements than the 20×20 array. The focal size decreased from $9.5 \text{ cm} \times 9.5 \text{ cm}$ to $5.2 \text{ cm} \times 5.2 \text{ cm}$,

and the area decreased to 30% of its original size. At 5 m, the 10×10 array focal spot exceeds the aperture size, while the 20×20 and 40×40 arrays produce focal spots of $48 \text{ cm} \times 48 \text{ cm}$ and $23.2 \text{ cm} \times 23.2 \text{ cm}$, respectively. The focal spot size of the 40×40 array decreases to 23% of that of the 20×20 array. We consider the near-field and far-field boundaries for antennas. The boundary distance is given by:

$$r_{\text{far}} \approx \frac{2D^2}{\lambda} \quad (15),$$

where D is the antenna aperture (in meters), λ is the wavelength (in meters), and r_{far} is the boundary distance between the far field and near field (in meters). Distances greater than this calculated value are far-field; distances less than this value are near-field. In the 5.8 GHz frequency band, the near-field range for the 20×20 array is less than 9.7 m, while that for the 40×40 array is less than 38.7 m. Based on the above data, within the 5 – 10 m near-field focusing range, the performance improvement of the 40×40 array over the 20×20 array is less than the increase in the number of units. Therefore, the 20×20 array is a more practical choice.

3 Design of 3-Bit Reflective RIS and Simulation

3.1 3-Bit Reflective RIS Design

Based on the theoretical analysis in Section 2.1 and the simulation results in Section 2.3, we designed a high-reflectivity, reconfigurable, focusing metasurface unit with a dog-bone-shaped patch for 5.8 GHz operation. Each unit consists of two dielectric substrate layers of different thicknesses and three copper metal patch layers of different shapes, which are bonded using prepreg materials. The patch is disconnected in the middle, and a varactor diode, specifically the MAVR-000120-1411 model produced by MACOM, is used to connect the two separated patches. The bottom layer consists of the DC bias line for the varactor diode and, through via connections, links to the top patches. The control terminal is connected via a feeding line, enabling adjustments to the input voltage at the control terminal, which

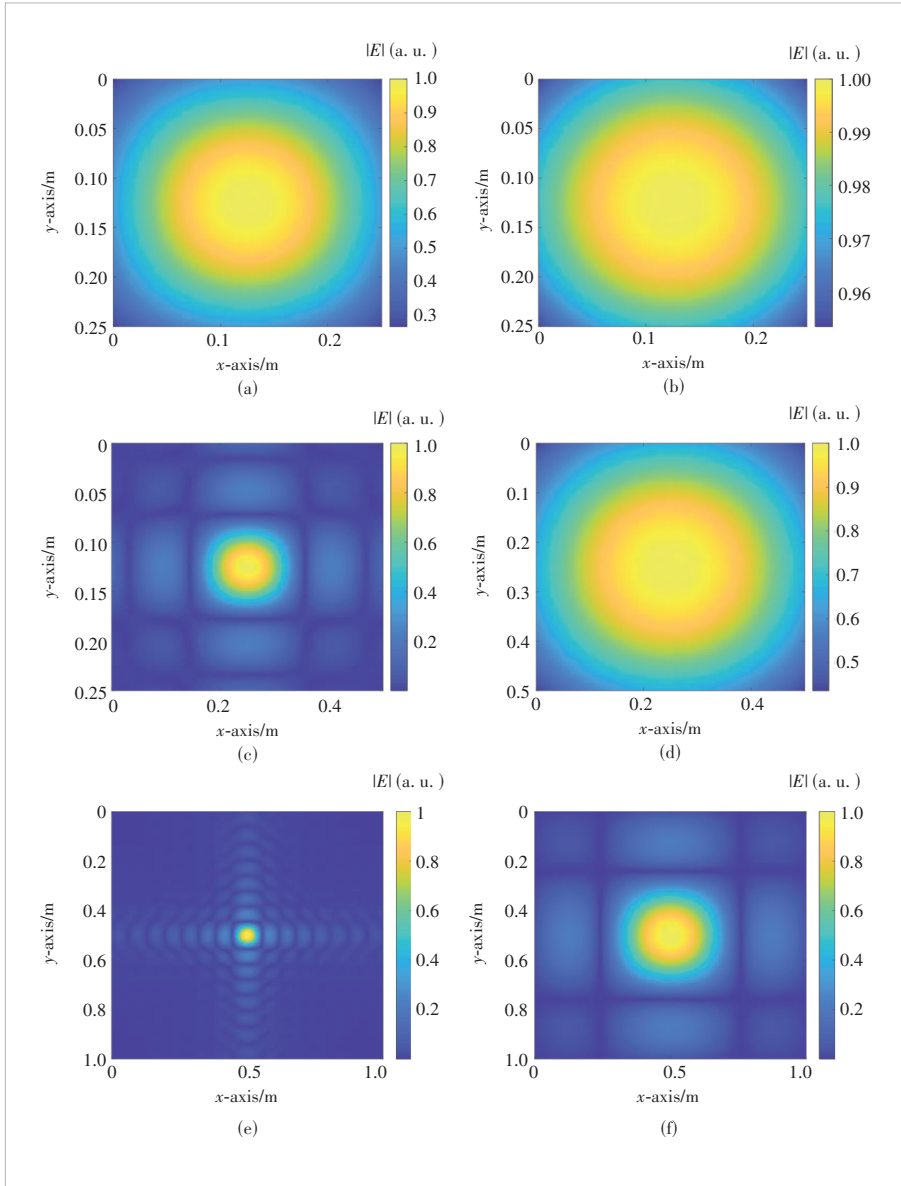


Figure 5. Normalized electric field distribution of different array sizes and focus distances: (a) 10×10 array at (0, 0, 1 m); (b) 10×10 array at (0, 0, 5 m); (c) 20×20 array at (0, 0, 1 m); (d) 20×20 array at (0, 0, 5 m); (e) 40×40 array at (0, 0, 1 m); (f) 40×40 array at (0, 0, 5 m)

ultimately changes the phase response of the unit. Fig. 6 shows the unit design. Table 1 lists the geometric parameters and material properties.

We used the full-wave electromagnetic simulation software ANSYS High-Frequency Structure Simulator (HFSS) to simulate the electromagnetic properties of the designed metasurface unit. For the varactor diode, three series lumped ports were used to model its resistor-inductor-capacitor (RLC) series equivalent circuit, as shown in Fig. 7, with boundary conditions assigned to its capacitance, resistance, and inductance. To determine the varactor diode parameters corresponding to 3-bit phase control, a Floquet port was employed for 2D

periodic S-parameter analysis. A parameter sweep was set up for the capacitance C_1 in the RLC circuit, scanning from 0.15 pF to 0.95 pF.

The 3-bit control corresponds to eight states, with the phase response of each unit differing by 45° between adjacent states. The total phase shift across the eight states must be at least 315° to cover a 360° phase range. The reflection phase and amplitude curves for the incident polarization are obtained from the S_{11} data. Figs. 8 and 9 show the parameter curves for the eight capacitance values. Table 2 lists the capacitance, encoding, bias voltage, and phase for each state.

From the S_{11} phase curves, the phase difference between adjacent states is approximately 45° , with the maximum difference exceeding 315° at 5.8 GHz. Due to phase deviations in practical systems, achieving perfect phase intervals is challenging. Additionally, the limited control capability of the diode makes it difficult to meet large phase differences across a wide frequency band. Therefore, a phase deviation criterion is introduced as the evaluation standard for multi-bit unit design.

To quantify the phase error versus frequency, the equivalent bit number has been employed in previous works for programmable metasurface reflectors with discrete phase steps. Pereira et al. proposed a quality indicator for the phase response of discretely tunable reflect arrays as the equivalent bit number^[30]. The equivalent bit number is defined as:

$$N_{\text{bit}} = \log_2 \left(\frac{360}{\sqrt{12} \sigma} \right) \quad (16),$$

where σ is the phase standard deviation (in degrees), calculated from the set of phase differences, $\Delta\varphi$, between different discrete states:

$$\sigma = \sqrt{\frac{\sum_{m=1}^M \Delta\varphi^3}{12 \times 360}} \quad (17).$$

The upper error limit for 3-bit control is 16.25° . After calculation, the standard phase deviation of the designed unit in

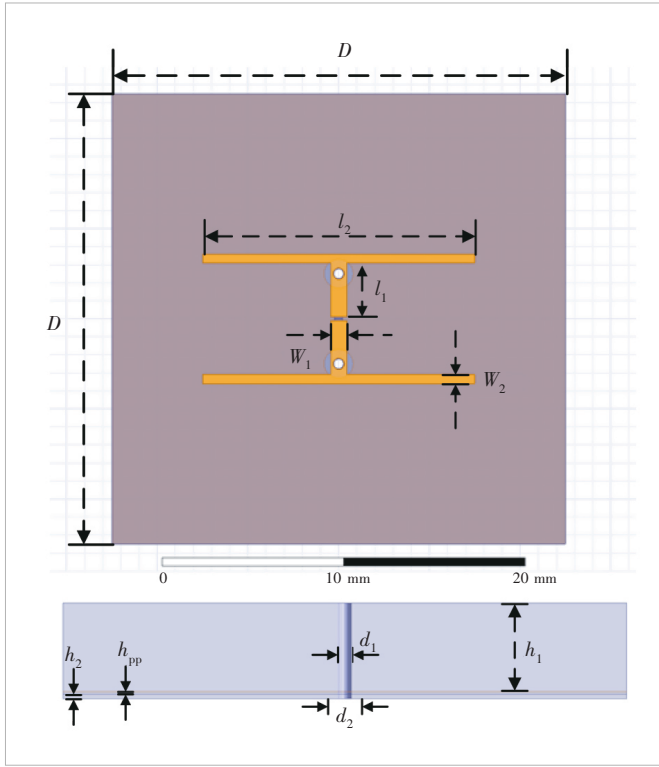


Figure 6. Design diagram of the unit

this study is 13.3° , which is below the upper limit, indicating that the design meets the 3-bit control requirements.

From the S_{11} amplitude curves, the reflection amplitudes for the eight modulation states are all greater than -3.3 dB within the operating frequency band, indicating low energy loss. The unit thus achieves phase control with nearly constant reflectivity. This is attributed to the dielectric substrate materials, whose appropriate dielectric constant enhances the reflection efficiency of the metasurface.

3.2 Simulation and Discussion

By periodically extending the unit model from Section 3.1 in two dimensions, we es-

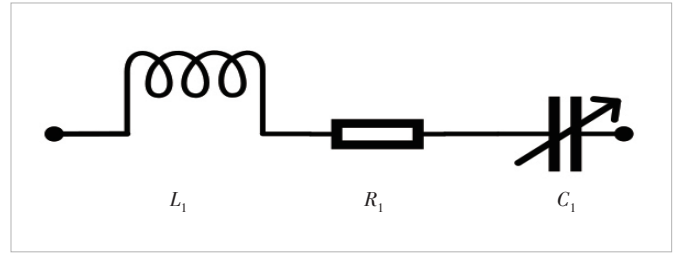


Figure 7. Equivalent circuit of the varactor diode, modeled as a series RLC circuit

tablished an array model in HFSS. Incident antenna feeds were used as inputs to investigate the focusing effect and efficiency of the array.

Microstrip antennas were used as feeds owing to their smaller volume, which can significantly reduce the obstruction of reflected focused beams. Because of their poor directivity, the microstrip antenna and the metasurface were placed at a location that ensures high feed efficiency, with the feed source at $(0, 0, 0.08)$ m. The incident power is 1 W. We used a 10×10 metasurface array and set the focusing position at the center of the reference plane, 1 m away from the center of the metasurface, to evaluate the theoretical focusing performance. The compensation phase of each element is calculated using the phase calculation formula in Section 2.2.

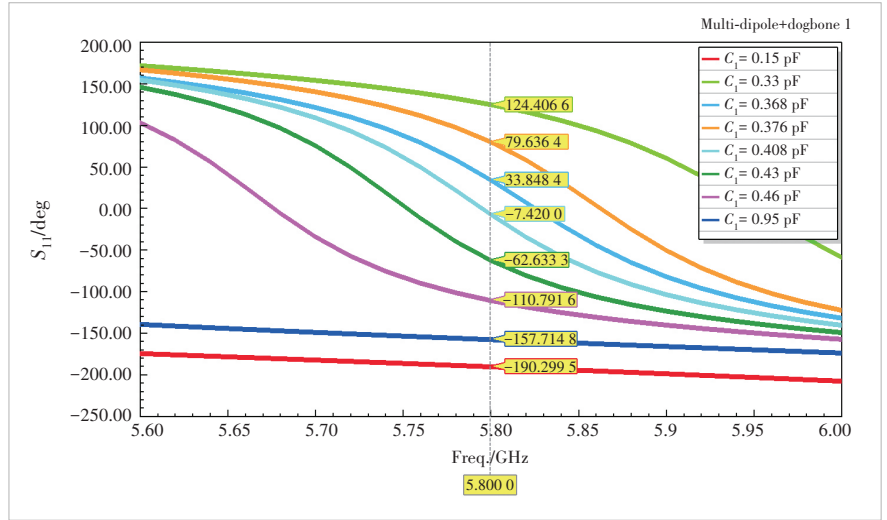
Figure 8. S_{11} phase curves for different capacitance values

Table 1. Geometric parameters and material properties of the unit design

Geometric Parameters	Element dimensions and thickness	$D = 25$ mm	$h_1 = 4$ mm
		$h_2 = 0.2$ mm	$h_{pp} = 0.1$ mm
	Patch parameters	$w_1 = 0.9$ mm	$l_1 = 2.985$ mm
		$w_2 = 0.5$ mm	$l_2 = 15$ mm
	Via dimensions	$d_1 = 0.6$ mm	$d_2 = 1.6$ mm
Material Properties	Dielectric substrate	F4B material (with a relative permittivity of 2.65 and a loss tangent of 0.001 5)	
	Adhesive material	Prepreg material (with a relative permittivity of 2.5 and a loss tangent of 0.004)	

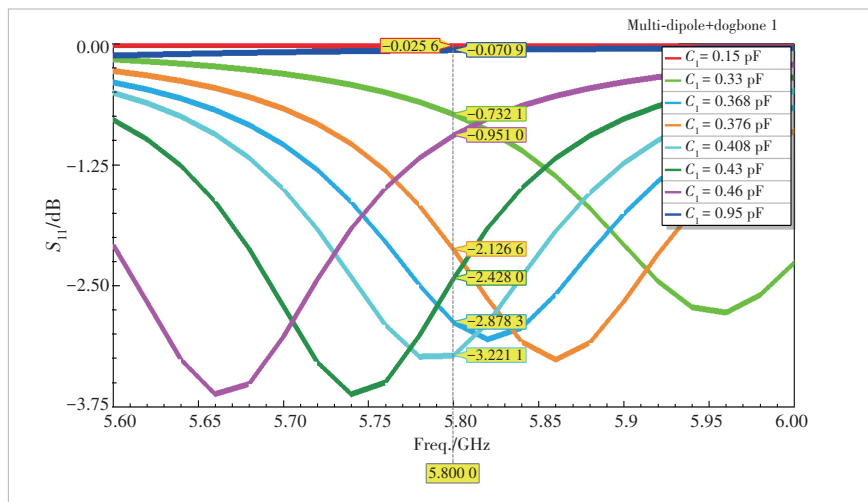


Figure 9. S_{11} amplitude curves for different capacitance values

Table 2. 3-bit encoding correspondence table

Encoded Value	Bias Voltage/V	Capacitance/pF	Resistance/ Ω	Inductance/nH	Phase/deg
000	9.60	0.150	0.88	0.02	124.41
001	4.56	0.330	0.88	0.02	79.64
010	4.08	0.368	0.88	0.02	33.85
011	3.70	0.376	0.88	0.02	-7.42
100	3.32	0.408	0.88	0.02	-62.63
101	2.94	0.430	0.88	0.02	-110.79
110	2.12	0.460	0.88	0.02	-157.71
111	0	0.950	0.88	0.02	-190.30

The phase compensation and focusing effect diagrams are shown in Fig. 10.

When the 10×10 array is focused at the center of the reference plane 1 m away, the size of the focused area is approximately $19 \text{ cm} \times 19 \text{ cm}$, which aligns with the simulation results in Section 2.3, verifying the accuracy of the WPT array focusing simulation. To determine the focusing efficiency of the array, the field calculator in HFSS software was used. We selected the focal area on the reference surface that receives energy from the metasurface as the integration area. The real part of the Poynting vector was integrated over this area. The total energy in the focal region is 0.11 W, while the transmission power of the microstrip antenna is 1 W. According to the efficiency formula, the focusing efficiency is 11%. We note that we did not use a horn antenna with better directivity to reduce antenna obstruction in the simulation.

Using a horn antenna as the feed would further improve the focusing efficiency.

For the RIS-based WPT system simulated in Fig. 10, we set reference surfaces at different distances and calculated the focusing power using the Poynting vector integral from the simulation. The beam focusing efficiency mentioned in Section 2.3 was calculated at each distance. For comparison, we also included a high-gain beam as a reference. As shown in Fig. 11, at the set focusing distance, the focusing efficiency exceeds 11%, but it decreases significantly as the distance deviates. This indicates that deviation in target device positioning can lead to substantial changes in focusing efficiency, which can be resolved with more accurate location information.

Compared with the high-gain beam, the focusing beam achieves significantly higher efficiency at the focus distance, while the high-gain beam maintains relatively uniform efficiency across distances. For long-distance transmission, high-gain beams have advantages. However, for the short-range energy transmission scenario discussed in this paper, near-field beam focusing is more suitable, demonstrating the feasibility of the proposed scheme.

4 WPT Experiment

Based on the metasurface elements discussed in Section 3.1 and the array size discussed in Section 2.3, we fabricated a 20×20 WPT RIS prototype. The RIS panel measures $500 \text{ mm} \times 574 \text{ mm}$, as shown in Fig. 12. The DC bias line for controlling the reverse bias voltage of the diode is guided to the back of the RIS via a through-hole, and the input voltage is adjusted by a controller with 20 sets of 21 pins per group. The 21×20

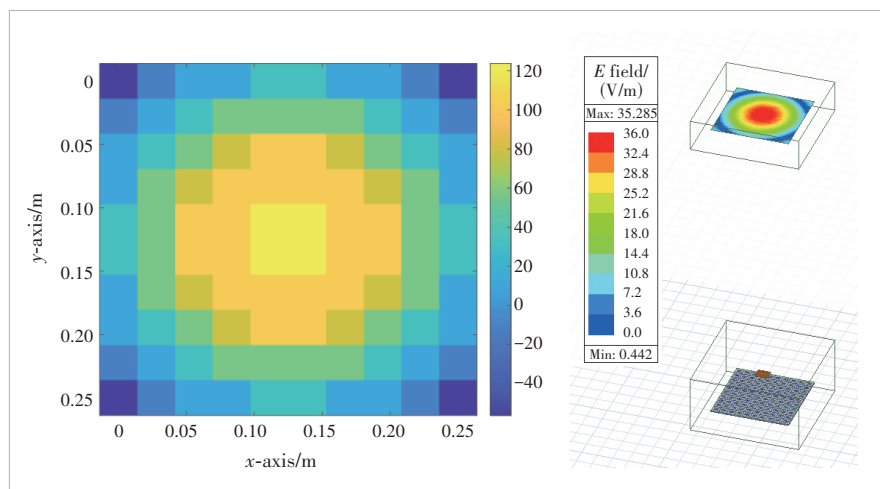


Figure 10. Phase compensation (left) and focusing effect (right) of the array

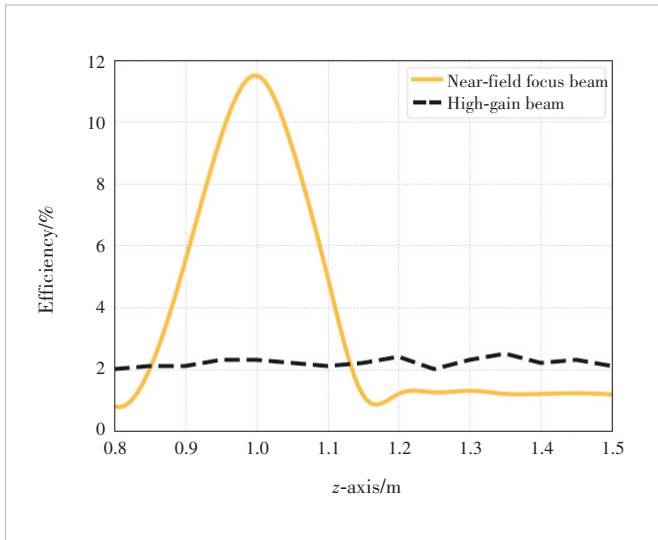


Figure 11. WPT efficiency versus distance for near-field focus beam and high-gain beam

(total of 420) bias lines are used because the 20×20 array has 400 units, with each column equipped with a grounding bias line to enhance grounding and system stability. The RIS is regulated by a high-performance FPGA control board, which can output 480 DC channels. The RIS is connected to the FPGA control board via a ribbon cable.

We measured the focusing effect of near-field RIS in a microwave anechoic chamber. The configuration of the measuring device is shown in Fig. 13. The RIS is fed by a horn antenna, which is set to oblique incidence at positions $(0, -0.8, 0.5)$ m and operates at a frequency of 1 – 18 GHz. Because near-field feed blockage is more severe than the far-field^[11], the focusing effect decreases significantly when the focusing position is behind the horn antenna. Therefore, oblique incidence is used for energy feeding. However, the oblique incidence results in nonuniform illumination of the RIS aperture, which reduces the focusing efficiency^[31].

The test used planar near-field scanning with a low-gain rectangular waveguide probe to scan a $50 \text{ cm} \times 50 \text{ cm}$ plane.

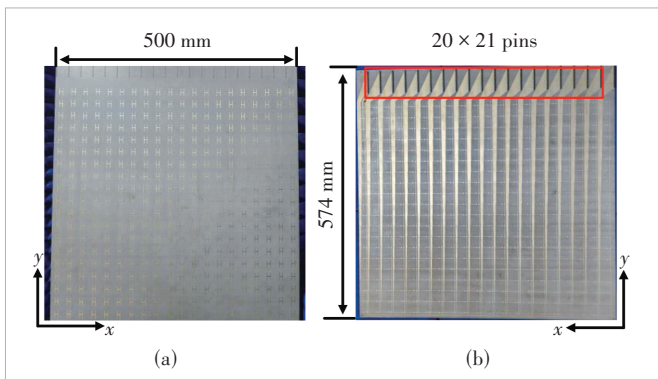


Figure 12. Fabricated 3-bit RIS: (a) front view; (b) back view

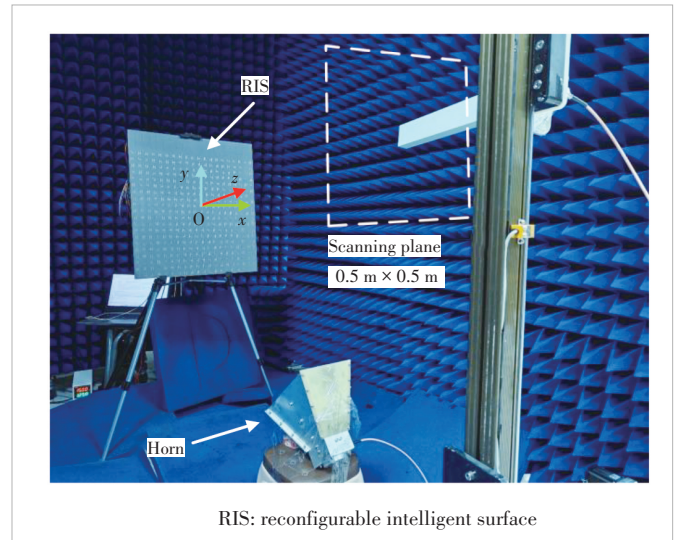


Figure 13. Near-field measurement setup for focal spots in the anechoic chamber

During the test, the polarization mode of the RIS and probe remained consistent. Due to site limitations, the focusing distances were set to 1 m and 3 m, with focusing positions at $(0, 0, z)$ m on the focal plane (where z is 1 m and 3 m) and at $(-0.1, -0.1, z)$ m, respectively. The results are shown in Fig. 14. For the convenience of observation, all electric field distribution results are normalized.

It can be observed that there are obvious focal points at the designed focal positions, which verifies the rationality of phase compensation for near-field focusing. The focal point contours in Fig. 14 appear elliptical, likely due to oblique incidence and the non-uniform dielectric constant of the F4B substrate used. As the distance increases, the focal point size grows significantly: at 1 m, it is $9.5 \text{ cm} \times 9.5 \text{ cm}$; at 3 m, it is $27 \text{ cm} \times 27 \text{ cm}$. The test results at a focusing distance of 1 m are consistent with the simulation results. For the 3 m case, the simulated focus size (with the configuration modified accordingly in Section 2.3) is $27.6 \text{ cm} \times 27.6 \text{ cm}$, and the measured results also match the simulation results. These test results indicate that the designed near-field focusing scheme and 3-bit RIS can focus wireless power at the desired position and remain effective at longer distances.

To verify the effectiveness of energy focusing, we used a directional microstrip antenna array as the RX and connected it to a VNA to measure the received power. The transmitting antenna configuration was the same as above. The microstrip antenna is a commercial antenna array measuring $14 \text{ cm} \times 14 \text{ cm}$, with a nominal gain of 20.5 dBi at 5.8 GHz. Placed at $(0, 0, z)$ (where z is 1 m and 3 m), the measurement setup and RX antenna are shown in Fig. 15. The energy transmission efficiency is determined by the ratio of the received power to the transmitted power, with the transmitted power fed into the horn antenna set at -5 dBm . Measurements show that the energy transmission efficiency is about 10% (-15 dBm) at a dis-

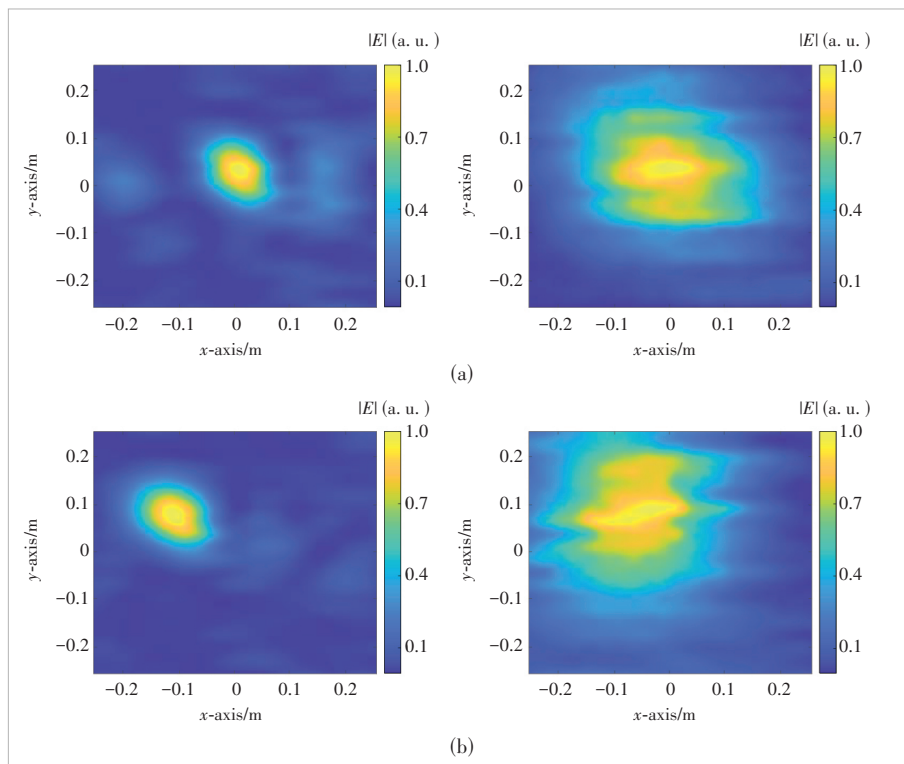


Figure 14. Measured normalized electric field distributions with focusing distance (represented as z) set to 1 m (left column) and 3 m (right column): (a) focusing position at $(0, 0, z)$ m; (b) focusing position at $(-0.1, -0.1, z)$ m

able part of daily life. In this study, we designed a near-field reflective focusing metasurface with a 3-bit quantization precision and high reflectivity. Simulations were conducted to investigate its specific energy focusing efficiency and the advantages provided by its reconfigurability. We explored the phase control mechanism of the metasurface, achieving beam focusing through phase compensation based on near-field beam coherent superposition theory. Analysis identified phase quantization precision and array aperture as the two primary factors affecting focusing performance and provided optimal parameter values. A metasurface unit based on a dog-bone-shaped patch structure was designed and simulated, using varactor diodes to achieve 360° non-quantized phase control with high reflectivity at 5.8 GHz. Actual measurements further validated the feasibility of the focusing strategy and unit design for near-field WPT.

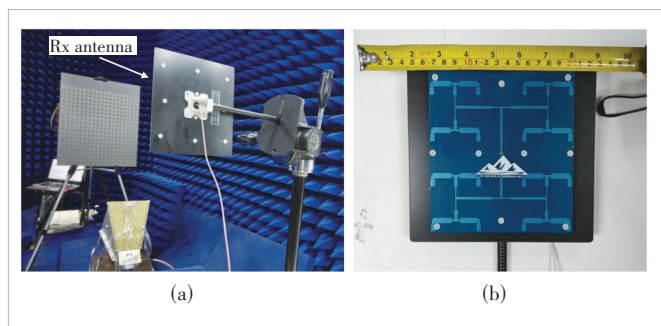


Figure 15. Measurement setup for WPT efficiency: (a) anechoic chamber setup; (b) Rx antenna

tance of 1 m and about 2.5% (-21 dBm) at a distance of 3 m. It is noted that the focusing efficiency significantly decreases at 3 m because the receiving area is fixed, while the 3 dB focal spot is significantly larger than the receiving end area, so most energy is not received. Compared with the 1-bit scheme in Ref. [11], the 3-bit scheme achieves more efficient near-field WPT at longer distances.

5 Conclusions

The development of wireless power transfer technology has spanned a century, and it is foreseeable that it will be widely adopted in the near future, becoming an indispens-

References

- [1] Ponnimbaduge Perera T D, Jayakody D N K, Sharma S K, et al. Simultaneous wireless information and power transfer (SWIPT): recent advances and future challenges [J]. *IEEE communications surveys & tutorials*, 2018, 20 (1): 264 – 302. DOI: 10.1109/COMST.2017.2783901
- [2] Valenta C R, Durgin G D. Harvesting wireless power: survey of energy-harvester conversion efficiency in far-field, wireless power transfer systems [J]. *IEEE microwave magazine*, 2014, 15(4): 108 – 120. DOI: 10.1109/MMM.2014.2309499
- [3] Powercast Co. Wireless power FAQs [EB/OL]. (2024-02-01)[2024-07-16]. <https://www.powercastco.com/faq>
- [4] Powercast Co. Powercast's wireless power calculator [EB/OL]. (2024-02-01)[2024-07-10]. <https://calculator.powercastco.com>
- [5] Ossia. How does wireless charging work? [EB/OL]. (2023-08-23)[2024-07-16]. <https://www.ossia.com/blog/how-does-wireless-charging-work>
- [6] Xiaomi. Mi air charge technology [EB/OL]. (2021-04-20) [2024-12-07]. <https://www.mi.com/global/discover/article?id=1653>
- [7] Xue H, Wu X N, Cui X W, et al. Wireless power transfer based on Bessel beam [C]//*Proceedings of IEEE International Workshop on Electromagnetics: Applications and Student Innovation Competition (iWEM)*. IEEE, 2021: 1 – 3. DOI: 10.1109/iWEM53379.2021.9790507
- [8] Xue H, Wu X N, Cui X W, et al. Multitarget wireless power transfer system using metasurface for quasi-Bessel beams with large half power beam length [J]. *IEEE transactions on microwave theory and techniques*, 2022, 70(10): 4449 – 4462. DOI: 10.1109/TMTT.2022.3197619
- [9] Wu Q Q, Zhang R. Weighted sum power maximization for intelligent reflecting surface aided SWIPT [J]. *IEEE wireless communications letters*, 2020, 9(5): 586 – 590. DOI: 10.1109/LWC.2019.2961656
- [10] Li L, Zhang P, Cheng F J, et al. An optically transparent near-field focusing metasurface [J]. *IEEE transactions on microwave theory and techniques*, 2021, 69(4): 2015 – 2027. DOI: 10.1109/TMTT.2021.3061475

- [11] Han J Q, Li L, Ma X J, et al. Adaptively smart wireless power transfer using 2-bit programmable metasurface [J]. *IEEE transactions on industrial electronics*, 2022, 69(8): 8524 – 8534. DOI: 10.1109/TIE.2021.3105988
- [12] Tran N M, Amri M M, Park J H, et al. Multifocus techniques for reconfigurable intelligent surface-aided wireless power transfer: theory to experiment [J]. *IEEE Internet of Things journal*, 2022, 9(18): 17157 – 17171. DOI: 10.1109/JIOT.2022.3195948
- [13] Zhang P, Li L, Zhang X M, et al. Design, measurement and analysis of near-field focusing reflective metasurface for dual-polarization and multi-focus wireless power transfer [J]. *IEEE access*, 2019, 7: 110387 – 110399. DOI: 10.1109/ACCESS.2019.2934135
- [14] Dai L L, Wang B C, Wang M, et al. Reconfigurable intelligent surface-based wireless communications: antenna design, prototyping, and experimental results [J]. *IEEE access*, 2020, 8: 45913 – 45923. DOI: 10.1109/ACCESS.2020.2977772
- [15] Xu H X, Tang S W, Ma S J, et al. Tunable microwave metasurfaces for high-performance operations: dispersion compensation and dynamical switch [J]. *Scientific reports*, 2016, 6: 38255. DOI: 10.1038/srep38255
- [16] Araghi A, Khalily M, Safaei M, et al. Reconfigurable intelligent surface (RIS) in the sub-6 GHz band: design, implementation, and real-world demonstration [J]. *IEEE access*, 2022, 10: 2646 – 2655. DOI: 10.1109/ACCESS.2022.3140278
- [17] Cui T J, Liu S, Zhang L. Information metamaterials and metasurfaces [J]. *Journal of materials chemistry C*, 2017, 5(15): 3644 – 3668. DOI: 10.1039/c7tc00548b
- [18] Ali A, Mitra A, Ayssa B. Metamaterials and metasurfaces: a review from the perspectives of materials, mechanisms and advanced metadevices [J]. *Nanomaterials*, 2022, 12(6): 1027. DOI: 10.3390/nano12061027
- [19] Wang J F, Li Y, Jiang Z H, et al. Metantenna: when metasurface meets antenna again [J]. *IEEE transactions on antennas and propagation*, 2020, 68(3): 1332 – 1347. DOI: 10.1109/TAP.2020.2969246
- [20] Faenzi M, Minatti G, González-Ovejero D, et al. Metasurface antennas: new models, applications and realizations [J]. *Scientific reports*, 2019, 9: 10178. DOI: 10.1038/s41598-019-46522-z
- [21] Ding F, Pors A, Bozhevolnyi S I. Gradient metasurfaces: a review of fundamentals and applications [J]. *Reports on progress in physics*, 2018, 81(2): 026401. DOI: 10.1088/1361-6633/aa8732
- [22] Lin J, Genevet P, Kats M A, et al. Nanostructured holograms for broadband manipulation of vector beams [J]. *Nano letters*, 2013, 13(9): 4269 – 4274. DOI: 10.1021/nl402039y
- [23] Fu X J, Yang F, Liu C X, et al. Terahertz beam steering technologies: from phased arrays to field-programmable metasurfaces [J]. *Advanced optical materials*, 2020, 8(3): 1900628. DOI: 10.1002/adom.201900628
- [24] Calà Lesina A, Goodwill D, Bernier E, et al. Tunable plasmonic metasurfaces for optical phased arrays [J]. *IEEE journal of selected topics in quantum electronics*, 2021, 27(1): 4700116. DOI: 10.1109/JSTQE.2020.2991386
- [25] Lindfors K, Dregely D, Lippitz M, et al. Imaging and steering unidirectional emission from nanoantenna array metasurfaces [J]. *ACS photonics*, 2016, 3(2): 286 – 292. DOI: 10.1021/acsphotonics.5b00646
- [26] Yi X J, Chen X, Zhou L, et al. A microwave power transmission experiment based on the near-field focused transmitter [J]. *IEEE antennas and wireless propagation letters*, 2019, 18(6): 1105 – 1108. DOI: 10.1109/LAWP.2019.2910200
- [27] Tran N M, Amri M M, Park J H, et al. Reconfigurable-intelligent-surface-aided wireless power transfer systems: analysis and implementation [J]. *IEEE Internet of Things journal*, 2022, 9(21): 21338 – 21356. DOI: 10.1109/JIOT.2022.3179691
- [28] Di Renzo M, Zappone A, Debbah M, et al. Smart radio environments empowered by reconfigurable intelligent surfaces: how it works, state of research, and the road ahead [J]. *IEEE journal on selected areas in communications*, 2020, 38(11): 2450 – 2525. DOI: 10.1109/JSAC.2020.3007211
- [29] So S, Mun J, Park J, et al. Revisiting the design strategies for metasurfaces: fundamental physics, optimization, and beyond [J]. *Advanced materials*, 2023, 35(43): 2206399. DOI: 10.1002/adma.202206399
- [30] Pereira R, Gillard R, Sauleau R, et al. Dual linearly-polarized unit-cells with nearly 2-bit resolution for reflectarray applications in X-band [J]. *IEEE transactions on antennas and propagation*, 2012, 60(12): 6042 – 6048. DOI: 10.1109/TAP.2012.2213051
- [31] Yang H H, Yang F, Xu S H, et al. A 1-bit 10 × 10 reconfigurable reflectarray antenna: design, optimization, and experiment [J]. *IEEE transactions on antennas and propagation*, 2016, 64(6): 2246 – 2254. DOI: 10.1109/TAP.2016.2550178

Biographies

Wang Yunfeng received his bachelor's degree in communication engineering from Beijing University of Posts and Telecommunications (BUPT), China in 2024. He is currently working toward a master's degree in communication and information engineering at the School of Artificial Intelligence, BUPT. His research interests include the emerging technologies of 5G wireless communication networks.

Xu Zhiyu received his PhD degree in electrical and electronic engineering from the Hong Kong University of Science and Technology (HKUST), China in 2002. He is currently a senior architect at ZTE Corporation. He has authored or co-authored more than 10 articles in international journals and conferences, and has filed more than 10 patents. He is the author of two technical books. His research interests include 6G wireless communication and signal processing.

Zhang Hongtao (htzhang@bupt.edu.cn) received his PhD degree in communication and information systems from Beijing University of Posts and Telecommunications (BUPT), China in 2008. He is currently a full professor at BUPT. He has authored or coauthored more than 100 articles in international journals and conferences, and has filed more than 50 patents. He is the author of ten technical books. He is a senior member of the IEEE. His research interests include 5G wireless communication and signal processing.

An Embedded Database System for YANG



Han Bing¹, Gao Jirun¹, Zhang Kaiqi^{1,2}, Wang Hongzhi¹,
Chen Hedui³

(1. Harbin Institute of Technology, Harbin 150001, China;

2. Chongqing Research Institute of Harbin Institute of Technology,
Chongqing 401135, China;

3. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202603010

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260826.1320.002.html>,
published online August 27, 2026

Manuscript received: 2024-09-19

Abstract: Yet Another Next Generation (YANG) is a data modeling language which is primarily designed to define the hierarchical structure of data used in network configuration management protocols such as NETCONF. However, traditional relational databases struggle to maintain optimal index performance for YANG data, and it is challenging to ensure data validity and transaction characteristics, particularly in resource-constrained environments such as network base stations. In this paper, we introduce the design and implementation of DB4YANG, an embedded database optimized for minimal resource usage while effectively handling and storing YANG model data. DB4YANG supports fast read and write operations on YANG data and ensures data integrity through continuous synchronization with optimistic locking. This approach mitigates the performance drawbacks that arise from non-native processing and format conversions commonly found in relational databases. Furthermore, this paper delves into transaction processing techniques tailored for YANG model data. These techniques are designed to ensure the reliability and consistency of data operations and strictly adhere to the Atomicity, Consistency, Isolation, and Durability (ACID) principles to preserve data integrity. We conducted several experiments comparing DB4YANG with an existing database, and the results demonstrate the effectiveness and efficiency of the proposed database.

Keywords: YANG; embedded database; high-performance index; data consistency

Citation (Format 1): Han B, Gao J R, Zhang K Q, et al. An embedded database system for YANG [J]. *ZTE Communications*, 2026, 24(3): 82 – 91. DOI: 10.12142/ZTECOM.202603010

Citation (Format 2): B. Han, J. R. Gao, K. Q. Zhang, et al., “An embedded database system for YANG,” *ZTE Communications*, vol. 24, no. 3, pp. 82 – 91, Sept. 2026. doi: 10.12142/ZTECOM.202603010.

1 Introduction

In recent years, the global expansion of cloud computing and the surge in business demands have accelerated network technologies and software-defined networking (SDN)^[1]. The essence of SDN is the decoupling of the control plane from the data plane, which was originally achieved via OpenFlow. While OpenFlow is no longer essential, the idea of separating control from forwarding remains central to SDN^[2]. The programmability of networks has progressively evolved into a crucial criterion for evaluating SDN frameworks.

Traditionally, programming network devices means using command line interfaces (CLI) and the simple network management protocol (SNMP). Network operators build scripting and management tools on top of these protocols to achieve some level of network programmability, which is useful for automating numerous tasks. Some devices allow for configuration through a web interface using extensible markup lan-

guage (XML) for bulk settings. However, despite its ease of use, CLI is not efficient or easy to read. It is primarily designed for human-machine interaction rather than the programmatic needs of networks. CLI proves insufficient for large-scale network automation due to its limited programmability and the complexities involved in development.

SNMP was introduced to simplify device configuration. It is a network management protocol comprising application-layer protocols, database schemas, and data objects for network supervision^[3]. SNMP is widely used for monitoring, but it suffers from several limitations. First, it lacks readability. It also has security vulnerabilities and lacks backup and rollback options. Moreover, its write capabilities are limited, which means it is primarily suitable for monitoring and basic configuration checks rather than in-depth configuration management. These limitations hinder its utility in network configuration tasks.

To address the intricacies of device configuration management, the IETF introduced standards including NETCONF, Yet Another Next Generation (YANG), and RESTCONF. NETCONF is an XML-based protocol for network management. It includes a suite of functionalities such as configuration, monitoring, fault management, secure authentication,

*The corresponding author is Zhang Kaiqi.

This work was supported by Natural Science Foundation of Chongqing, China, under Grant No. CSTB2023NSCQ-MSX0673 and ZTE Industry-University-Institute Cooperation Funds under Grant No. 2023ZTE09-02.

and access control. These features have contributed to its widespread industry adoption^[4]. YANG, defined in RFC 6020 (2010), is a modeling language for standardizing NETCONF's data models and operations. The integration of YANG models within NETCONF messages creates a universal network programmability interface. YANG-based network programmability has been widely adopted since 2016, simplifying network administration and reducing operational costs.

YANG provides a concise yet comprehensive means of defining data structures, yielding lightweight data descriptions. Network devices use remote procedure calls (RPCs) that are translated into XML messages, conforming to YANG model rules, to enable seamless end-to-end communication. However, with the advent of technologies such as 5G, managing the ever-growing volume of YANG model data at a large scale presents significant challenges.

Traditional relational databases have difficulty maintaining optimal index performance for YANG data, which necessitates separating model management from data management. Ensuring data integrity requires constant synchronization between these systems, which in turn introduces consistency challenges. Document databases such as MongoDB necessitate converting YANG data into the BSON format, which is a memory-intensive process and involves frequent data format conversions^[5]. These conversions can degrade performance, especially in resource-constrained settings like network base stations.

In such environments, devices produce vast amounts of configuration and state data, intricately detailed by YANG models. Typically, this data is stored directly in relational database management systems (RDBMSs). While RDBMSs are built for handling structured data, they face numerous inefficiencies when dealing with YANG-modeled data structures. Most RDBMSs lack native capabilities for managing YANG models and validating documents according to YANG schemas. Consequently, developers are compelled to devise supplementary modules to facilitate model governance and data validation, introducing an additional layer of complexity to system architectures and increasing maintenance overhead. This necessitates a reevaluation of conventional data storage strategies to accommodate the intricacies of YANG models more effectively and streamline system design for enhanced scalability and maintainability.

Moreover, most network devices cannot afford high levels of memory and computing resource consumption. Given the limited performance of these devices, they may struggle to handle complex data operations, run advanced algorithms, or manage a large number of network connections simultaneously. In this resource-constrained scenario, the performance bottleneck becomes particularly evident when dealing with massive network data or executing compute-intensive tasks, which could lead to degraded performance or even system crashes.

These problems motivate us to design and implement a data-

base that efficiently stores and indexes YANG model data and can be deployed across diverse environments without imposing substantial resource demands. We build DB4YANG on an efficient key-value storage structure with a high-performance indexing mechanism for parsing, storing, and managing YANG-modeled data, thereby simplifying both data management and retrieval. The data management module specializes in handling data objects constrained by YANG models, validating the correctness of XML or JSON documents, and converting them into corresponding objects for storage. Besides, to maintain data integrity and consistency, DB4YANG adopts optimistic locking to achieve concurrency control. The development of an embedded database for YANG models in low-storage-performance network transmission contexts is of considerable practical and industrial importance.

2 Related Work

2.1 Yang Model Application

YANG has emerged as the industry-preferred schema for data description, prompting standards bodies, vendors, operators, and over-the-top (OTT) service providers to define their respective YANG models. This trend has garnered attention from prestigious universities and research institutions worldwide, which are actively engaged in developing applications based on YANG models and researching ways to enhance network transmission performance through YANG-driven methodologies.

Golmohammadi^[6] developed Trace, a centralized tracing tool that triggers traceroute commands with a distinct syntax on transmission control protocol/Internet protocol (TCP/IP) routers from a network management system. It tracked network hops and BGP autonomous systems (ASes), measured round-trip time (RTT) to specific endpoints, and sent the data back to the management system. The authors also evaluated the viability of integrating this tracing tool with YANG-based network management systems.

Lee et al.^[7] improved the NETCONF protocol using YANG XML documents. They devised a method for creating data blocks based on device configuration dependencies, allowing for the logical organization of configuration data. This data can be stored and managed in separate segments, enhancing efficiency when multiple administrators work on different data blocks simultaneously, which makes this enhanced NETCONF outperform existing methods.

Dallaglio et al.^[8] proposed a control and management plane for elastic optical networks (EONs) using NETCONF and YANG. They introduced a YANG model for optoelectronic transponders supporting sliceability, variable bit rates, and monitoring. Their experimental validation confirmed the model's effectiveness in controlling and managing EONs.

Amlou et al.^[9] created a microservice-based architecture for network automation, utilizing OpenConfig YANG models and NETCONF. This architecture simplified network configura-

tion, automated tasks, and reduced errors. Their work demonstrated how this approach streamlined configuration and data collection, effectively managing complex networks. Performance assessments confirmed its robustness in handling detailed configurations and its fault tolerance.

Liu et al.^[10] introduced MAYA, a system that aligns YANG models from various vendors or organizations by examining node attributes. MAYA uses semantic similarity techniques to measure the distance between attributes within nodes from different YANG models. They developed a YANG model alignment algorithm based on the stable marriage problem (SMP) to create inter-model mappings. Case studies and experiments showed that MAYA meets the alignment needs of YANG models in real-world settings.

In NETCONF- and YANG-based implementations of the data plane, validation efforts typically focus on YANG model verification. However, gaps in validation can lead to configuration errors, such as incorrect IP addresses or ports. Martinovic et al.^[11] proposed enhancements to data model validation by comparing actual configurations against expected ones, using XML comparison techniques.

Overall, current research on YANG models is largely centered on advancing YANG-based network management technologies and refining YANG's implementation and performance within modern network architectures^[12]. Despite this, there are still significant gaps in the direct governance and management of YANG model data.

2.2 Embedded System

In embedded systems, database operations are typically subject to strict timing constraints. Within this ecosystem, a class of database solutions exists that can be integrated into applications as library functions. These are known as embedded database management systems (EDBMS) or more commonly as embedded databases^[13]. By eliminating the overhead of client-server architectures, these systems offer efficient storage directly within applications and often act as local storage solutions. Noted for their minimal resource consumption, embedded databases are dedicated solely to their host processes and are capable of storing data volumes up to the terabyte scale. They are highly scalable, customizable, and strongly compatible across various platforms. Moreover, they are designed to be resilient in the face of failures.

Since embedded databases function on the application side without a dedicated maintenance team, they are vulnerable to user errors and hardware malfunctions^[14]. This necessitates the implementation of robust fault-handling strategies. Well-known examples of embedded databases include SQLite^[15], LevelDB^[16], and RocksDB^[17].

The primary difference between embedded and non-embedded databases lies in their operational models. Unlike standalone database servers that operate as separate daemons, embedded databases share the same process space as the ap-

plication. It is important to note that simply having a database on an embedded device does not make it an embedded database; such databases are more accurately described as embedded mobile databases. While it is theoretically possible for embedded devices to run server-side programs for network databases, the term "embedded databases" refers specifically to databases^[18] that run locally, integrated with the application without the need for a separate server instance, and are bound to the application's life cycle.

In contrast, conventional database access models employ intermediaries such as JDBC or ODBC^[19], through which clients communicate with remote database servers responsible for actual data manipulations. As depicted in Fig. 1, this model adheres to the classical client-server architecture, where clients and servers, as discrete processes, can collaborate closely whether they reside on the same machine or span different computers and networks.

Sysrepo is an open-source application that supports the required abstraction^[20]. To the best of our knowledge, Sysrepo is the most widely used database for managing and storing YANG-related data. It maintains a datastore, a core component of a NETCONF-enabled device. This datastore contains the configuration state of the device as defined by YANG data models installed within it. The code was modified to function as a plugin-style application for Sysrepo. This modification introduced a Sysrepo Client Library within the application, enabling it to subscribe to YANG modules in Sysrepo's datastore, with protocol buffers serving as the interface between the client library and the datastore engine. When a new NETCONF configuration utilizing a subscribed YANG data model is received, this setup triggers the execution of the C application.

Netopeer2, the third software system working in tandem with Sysrepo, functions as a NETCONF server^[21]. Operating as a background service (or daemon) on the switch, it monitors the management network for incoming NETCONF connections. Remote devices configure the switch via NETCONF by interacting with the Netopeer2 service. Upon receiving a NETCONF configuration update from a remote device, Netopeer2 processes the data payload and validates it against the

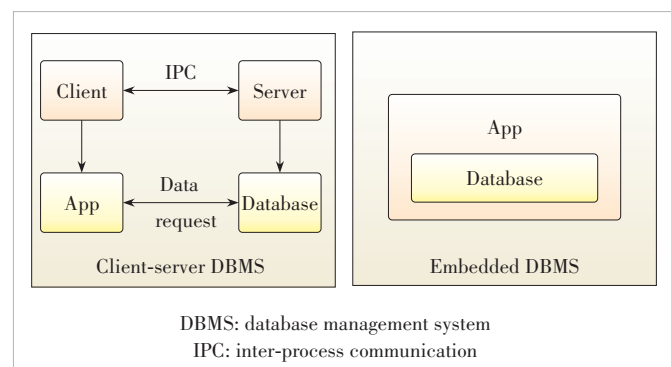


Figure 1. Architecture of client-server database and embedded database

YANG modules in the datastore to ensure consistency. After successful validation, the updated configuration settings are applied by invoking the relevant applications subscribed to the respective YANG model through Sysrepo.

3 Proposed Embedded Database Overview

3.1 Database Structure

The YANG model delegates functions to discrete modules, featuring an embedded design with minimal coupling. The architecture of the YANG model database facilitates client interaction through NETCONF and SNMP interfaces supported by the runtime environment. These interfaces convey YANG models and data to a database service centered around a data storage module. This service is meticulously designed to administer both model management and data management tasks.

The model management module, a core component, directly oversees the handling of YANG model files. Its responsibilities encompass the import/export, deletion, and modification of models, thereby ensuring the integrity and flexibility of the schema definitions. The data management subsystem is tasked with the storage of documents conforming to YANG model constraints within the database and adopts an optimized schema. It further implements specific routines for constraint validation and for performing create, read, update, and delete (CRUD) operations at the node level, enhancing data integrity and operability.

Presently, the model management and data management modules operate independently. These modules perform periodic synchronization of their datasets to maintain coherence. To augment functionality, an additional plugin subsystem that endows the system with data subscription capabilities has been integrated into the framework. Leveraging this subscription mechanism, affiliated business services can promptly react to modifications in the model data, thereby enabling real-time responsiveness.

Fig. 2 shows that the architecture of DB4YANG is constructed with different modules, each dedicated to a distinct function. It comprises the storage service module, the data management module, and the model management module, all of which rely on the B+ tree. Collectively, these modules form the backbone that primarily facilitates advanced model governance, robust data handling, and dynamic data subscription functionalities. Thereby the proposed database is a sophisticated and adaptable database infrastructure tailored for YANG-based network configurations.

3.2 Schema Module

DB4YANG employs a stratified approach to managing data. This system abstracts the minimal ensemble of YANG model files capable of describing a category of network devices into a “Container” for that category of devices. As YANG model files undergo parsing and type transformation, this Container

accumulates a multitude of “Types”, which may originate from typedef, container, or grouping nodes within the YANG schema. Guided by the dependencies among these types, the system constructs an intricate type hierarchy that serves as the basis for device description.

Acting as an abstraction of devices, a Container can be instantiated as a “Device”, representing a concrete, real-world network device comprising its entire informational footprint. Similarly, XML/JSON documents can be parsed and transformed into objects of a corresponding type, reflecting the concept of type instantiation. However, not all types possess the capability to interpret such documents. This capability is typically attributed to types derived from nodes such as container and RPC, highlighting their distinctive role in data interpretation and object creation.

The process of schema modification requires a base data container B and a schema modification sequence Q as input, where Q is a series of batch schema modification operations. Typical operations include inserting a new YANG file and removing constraints from a YANG file.

The schema modification process is divided into three phases: preparation, execution, and cleanup.

1) Preparation phase: Check if there is an ongoing schema modification on the base data container B ; if so, halt the process. Otherwise, create a backup of B 's model, retrieve all YANG files from B , and apply the schema modification sequence to generate a new set S of YANG files. Finally, create an execution thread.

2) Execution phase: Create a new data container C based on B 's data store, and attempt to add the YANG files from set S to C sequentially. If all additions succeed, the execution is

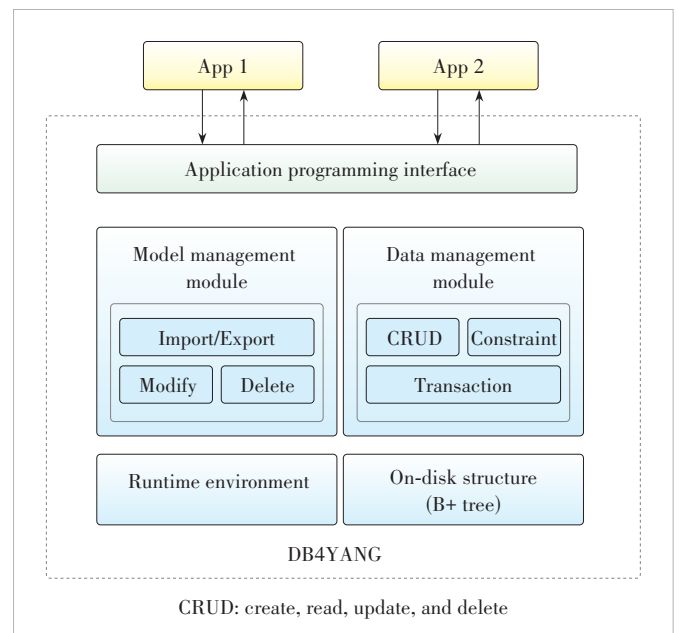


Figure 2. DB4YANG architecture

considered successful; if any addition fails, delete C .

3) Cleanup phase: If the execution succeeds, replace B with C , and update the metadata and YANG file set in external storage. If the execution fails, clean up the newly generated set S of YANG files.

3.3 Data Module

Compared with other databases, DB4YANG is an XML/JSON document database that operates under YANG schema constraints. It is required that specific data management operations be executed only on data containers where a schema tree definition has been established. In other words, any data management operation performed on a data container that has not undergone a model insertion operation will not succeed.

DB4YANG encodes documents that have passed validation into DB4YANG tuples and stores them in a B+ tree. The B+ tree is an N-ary sorted tree in external memory, with insertion and modification operations exhibiting relatively stable logarithmic time complexity. A typical B+ tree provides common key-value query interfaces such as GET, PUT, and RANGE. To adapt to the form of the B+ tree interface, DB4YANG assigns a unique identifier as a key to each piece of data before insertion, with the DB4YANG tuple serving as the value. The logical structure of the DB4YANG tuple is as follows: TXID, CTR_VER, DFLAG, and VER_PTR are all fixed-length 8-byte integers (stored in little-endian format). These fields are collectively used for the implementation of a multi-version concurrency control mechanism and document re-validation; DATA represents the actual stored document, encoded in the length and value (LV) format, with the length field encoded as a fixed-length 4-byte integer.

Furthermore, we propose a high-performance indexing design scheme that draws on the principles of inverted indexing (as illustrated in Fig. 4). This paper constructs a reverse mapping structure that maps query paths (QPATH) to object identifiers (OIDs), specifically HANDLE (analogous to document IDs in inverted indexes), thereby effectively simulating the fast query logic of inverted indexes. In particular, this paper introduces bitmap technology to further optimize HANDLE storage, which improves space efficiency.

Serving as an abstraction layer between users/applications and the underlying data storage, the external schema provides a concise and intuitive interface for data access. It is represented by a correspondence between a QPATH and an INDEX_PATH. Specifically, the QPATH is an XPath expression used to locate specific YANG model elements, where XPath serves as the standard notation for expressing query paths. The INDEX_PATH represents an internal identifier that points to a dedicated B+ tree. For each INDEX_PATH, a set of OIDs is persistently stored in a B+ tree

structure, with these OIDs corresponding to all documents matched by the XPath query.

From the perspective of the data management module, regardless of how complex the internal structure of the node-level index may be, its external representation remains a mapping of XPath to the B+ tree storing OIDs. This design allows external users or systems to access and manipulate indexes in a unified and straightforward manner, without concerning themselves with the underlying implementation details.

As shown in Fig. 5, DB4YANG adopts a refined and efficient metadata structure management strategy, centered on the use of a hash table to organize and access index metadata. The metadata is further stored in a B+ tree structure whose file path is /tmp/indexes/01.bp. This B+ tree not only ensures efficient data retrieval speeds but also carries key information about the index structures corresponding to each QPATH. Currently, the critical information in each metadata record is the actual storage location of the index (INDEX_PATH), which indicates the centralized storage point for all related HANDLES (existing in an OID form). To efficiently utilize storage space and accelerate query processing, this paper introduces an innovative interval partitioning scheme.

Considering that a HANDLE (OID) is a 64-bit integer,

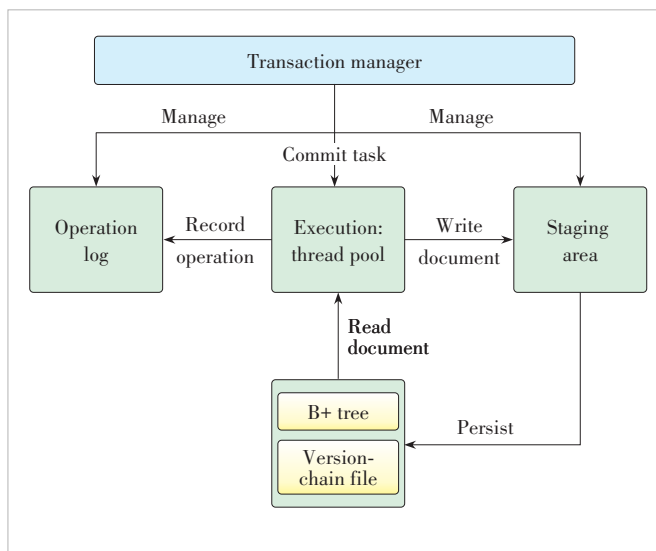


Figure 3. Structure of data management module

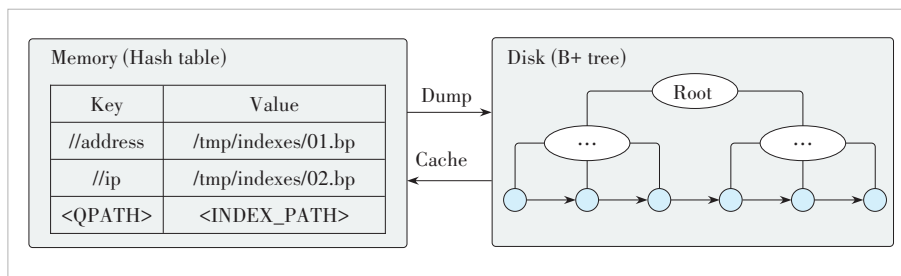


Figure 4. External view of index storage

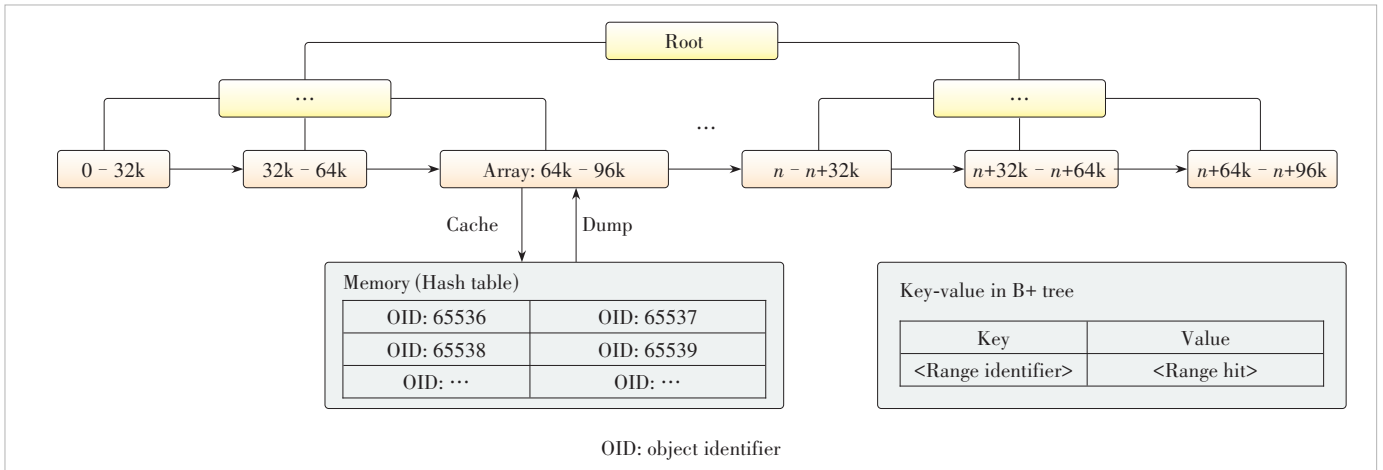


Figure 5. Internal view of index storage

this paper divides its numerical range into 2^{49} fixed-length subintervals, each of which corresponds to a fixed-size page of 4 096 bytes (i. e., one page), and uses the smallest HANDLE value within the interval as the unique interval identifier. This partitioning method not only simplifies the logic of index lookup but also improves the efficiency of memory and disk I/O operations through the preset fixed-size pages.

To store a HANDLE with a specific value of h , the system first determines the interval to which h belongs. Then, using bitmap technology, the system encodes and stores the set of HANDLES within the interval, optimizing storage space usage and enabling fast bitwise operation queries. In the bitset, each bit corresponds to a HANDLE, indicating its presence or absence, which greatly compresses the data volume. It also allows HANDLES to be inserted, deleted, and queried directly through bitwise operations, thereby enhancing processing speed.

4 Transaction Management

The YANG modeling language is designed for network device configuration, and most of its data definitions and operations are aligned with network management protocols such as NETCONF and RESTCONF. These include configurations, state data, remote procedure calls, and notifications. For network configurators, the process of setting up a device typically involves obtaining the corresponding YANG module and issuing configurations according to its schema. By contrast, developers might be tasked with authoring YANG modules, which requires adherence to YANG's syntax rules. Such tasks often necessitate meticulous drafting and verification by network operations personnel to ensure a high degree of operational accuracy in real-world deployment and application, thereby maintaining operational efficiency.

In practice, during the configuration of network devices, environments with high concurrency and intense data contention

are rarely encountered. Common application scenarios involve an individual network administrator batch-updating multiple network devices with configuration information or subscribing to device alerts. These operations typically involve minimal data conflicts, incur low retry costs, and require no absolute data consistency. Under such working conditions, optimistic locking is a more suitable choice than pessimistic locking for implementing concurrency control and transaction management.

A transaction is the unit of concurrency control in DB4YANG, defined as a sequence of operations issued by an application, and its execution satisfies the atomicity, consistency, isolation, and durability (ACID) properties. Fig. 6 shows that the DB4YANG data management module consists of components including a transaction manager, thread pool, staging area, operation log, B+ tree, and commit log, with the transaction manager at its core. DB4YANG provides a global default transaction manager while also allowing applications to use custom transaction managers. The set of transactions managed by a transaction manager is called the transaction group of that manager, and different transaction groups should be isolated from one another to guarantee inter-group transaction isolation. Concurrent transactions within a transaction group share the same thread pool, operation log, B+ tree, and commit log, while each transaction maintains its own staging area to stage the document modifications made by that transaction.

The operation log records significant actions during transaction execution processes, including the initiation, rollback, and commit of transactions, as well as the insertion, update, and deletion of documents. It is utilized for fault recovery upon a crash and subsequent reboot of DB4YANG. Based on the transaction execution characteristics of DB4YANG, the operation log is designed as a mixed UNDO/REDO log. Each operation log entry is encoded in binary for storage and comprises six fields: the transaction ID, operation type, container name, OID, old value, and new value.

When a transaction is selected by the scheduler, it transi-

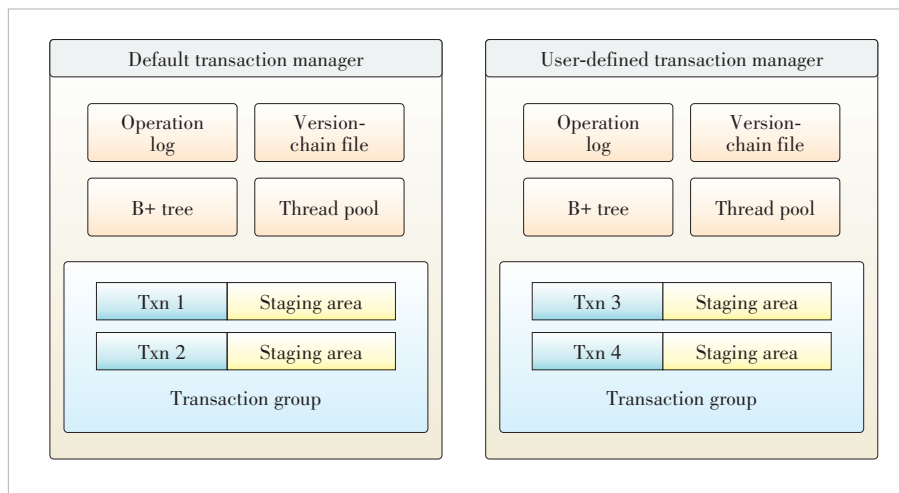


Figure 6. Transaction manager and transaction group

tions from the ready state to the running state. During this period, the transaction begins executing its database operations, such as queries, inserts, updates, or deletions. The system ensures the atomicity of operations within a single transaction, meaning that all operations either succeed entirely or fail altogether.

Once all operations within a transaction have been successfully executed, the transaction enters the pre-commit phase, a critical checkpoint for safety. Here, the transaction manager verifies the results of the transaction's execution and the current state of the system, ensuring that the transaction meets all predefined conditions and constraints (such as foreign key constraints and uniqueness constraints). This step is crucial for maintaining the consistency of the database.

If the pre-commit check succeeds, the transaction is formally committed. The system updates the commit log to record the transaction's success, a step that is key to achieving durability. Following this, the database modifications associated with the transaction have been permanently persisted, which marks the end of the transaction.

In contrast, if the commit check fails due to business rule violations or the detection of potential conflicts, the transaction is aborted. The system must then perform a rollback operation to revert the database state to its condition prior to the start of the transaction, ensuring that data consistency remains intact. In this process, all executed operations are recorded in the operation log, providing the necessary information for any subsequent retry or recovery.

In the event of unexpected exceptions during the transaction execution phase, such as system errors or insufficient resources, the system immediately triggers the rollback process of the transaction. All completed operations are rolled back sequentially, and the redo log and undo log (if applicable) are updated accordingly to record these changes and ensure the atomicity and durability characteristics of the transaction.

The data affected by the transaction is thereby restored to its initial state, as if the transaction had never been executed.

This transaction processing flow demonstrates how, through stage delineation and state management, data reliability and transaction-handling efficiency can be ensured amid complex and dynamic database operations.

5 Experiment

Sysrepo is a YANG-based configuration and operational state data store for Linux applications. Applications can use Sysrepo to store their configuration modeled by a provided YANG model instead of relying on flat configuration files, for example. Sysrepo ensures the consistency of the data stored in the

datastore and enforces the data constraints defined by the YANG model. Sysrepo can be easily integrated with management agents such as NETCONF or RESTCONF servers, using the same client library API that applications use to access their configuration. To our knowledge, Sysrepo is widely used to manage and store YANG-related data.

In this section, we verify system performance by comparing each operation with its counterpart in Sysrepo. We evaluate the schema modification, document insertion, and document reading performance of DB4YANG through automated programs, analyze the performance issues present in the current version, and discuss relevant improvement strategies.

To evaluate the performance of our database and Sysrepo, we conducted comparative experiments on a workstation equipped with an AMD Ryzen 7 4800H processor (8 cores, 4.2 GHz), 16 GB of DDR4 DRAM, and 512 GB of SSD storage.

In Figs. 7 and 8, the performance of the proposed database significantly surpasses that of Sysrepo for the following reasons. In Sysrepo, the verification of content changes is implemented through string comparison (see the function `lyd_anydata_equal`). When the configuration is large, the generated strings become lengthy, making comparisons time-consuming.

Configurations of subordinate modules that share the same top-level module are not segregated into separate files; instead, they are stored together in the top-level module's file. Consequently, when the data volume of one subordinate module grows large, loading this shared file becomes slow, which also degrades the performance of other modules.

Every modification and save operation involves rewriting the file. I/O operations are relatively slow, and the larger the data volume, the longer the time required. In addition, the shared memory file lock means that as long as one process is reading a large configuration, the lock is held, stalling other processes. YANG data is structured as a tree, in which the children of

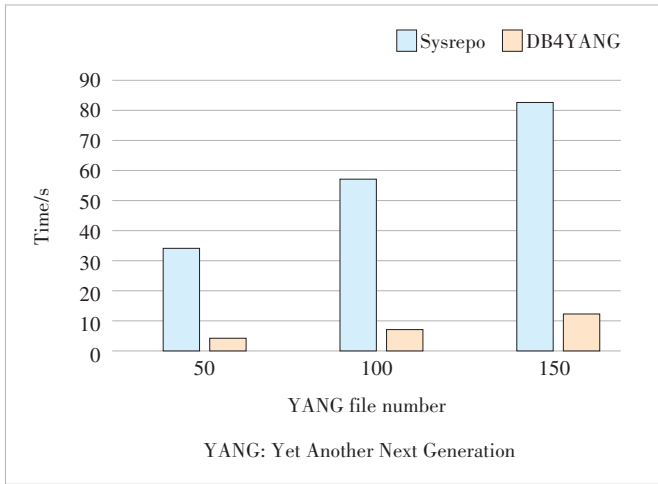


Figure 7. YANG schema import efficiency comparison

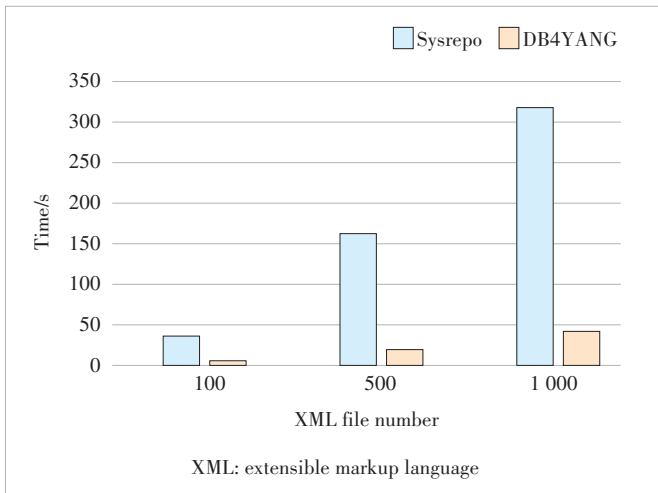


Figure 8. YANG document writing efficiency comparison

each node are organized as a doubly linked list. Since traversing a linked list requires linear time, this organization increases the latency of data retrieval and modification.

To optimize the performance of DB4YANG during system configuration, we load data for the requested module and its dependent modules for validation based on the configuration request. Due to interdependencies among configurations, any changes may impact related configurations, which must be considered during operations.

In the configuration-handling logic, changes must be synchronized across all submodules when operations such as deleting or replacing a top-level module's configuration are performed. This necessitates loading the data of all modules under the top-level module to ensure the correctness and consistency of the configuration. This step is essential to guarantee the accuracy of the configuration.

DB4YANG's caching functionality has also been opti-

mized. Previously, the entire data forest of the running datastore was cached as a single copy; in the optimized version, the cache is modularized. For the top-level modules of the project, we only cache their own data. For other top-level modules and their sub-modules, which are accessed less frequently, the modular splitting approach is not employed, and their data is cached as a whole. In this way, the modular caching strategy ensures efficient data management while reducing the complexity of cache handling.

To verify the transactional capabilities of our proposed database, we conducted an experiment, the results of which are depicted in Fig. 9. When the number of transactions reached 1 500, we executed a series of operations 100 times, comprising three transactional insertions, one point query, one full query, four XPath document-level queries, and four XPath node-level queries. Additionally, we performed one point-update operation and one point-deletion operation, each followed by a verification query to confirm that the update or deletion had taken effect. In total, these 15 operations constituted a single transaction. Following this pattern, when the number of transactions reached 15 000, the same set of operations was repeated 1 000 times; when it reached 150 000, the operations were executed 10 000 times.

Fig. 9 shows the influence of transaction numbers. Repeated experiments revealed that when all operations were XPath document-level queries, the transaction execution speed dropped to half of the average. Furthermore, in the absence of data deletion, queries gradually slowed down as data accumulated. As the data volume grows, querying increasingly becomes the critical factor limiting performance.

The main thread completes the model loading process, which does not become a bottleneck for the system. Concurrent read transactions are handled by child threads. And the functions `lyd_parse_data_mem` and `lyd_free_all` constitute the main bottleneck of the system.

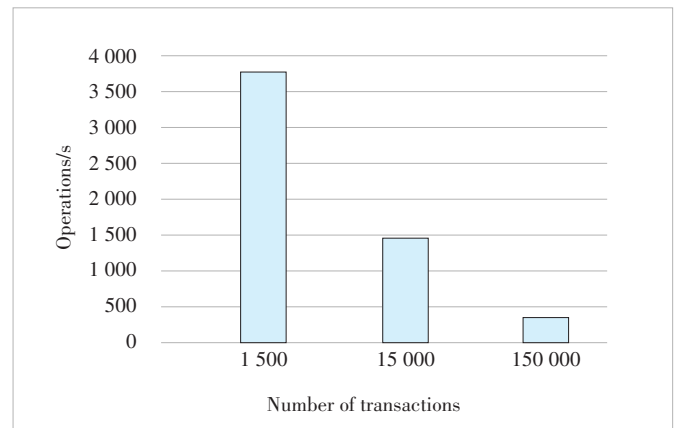


Figure 9. Performance of complicated transactional operations

Perf¹ is a performance analysis tool for Linux that can perform function-level and instruction-level hotspot detection. It was used to analyze the CPU usage of hotspot functions in programs, thereby locating performance bottlenecks.

Fig. 10 shows that in DB4YANG, the program spent 60.48% of its time loading YANG schemas and 22.82% on document insertion and reading. However, in Sysrepo, the majority of the program's runtime was consumed by loading YANG schemas. Due to the relatively small data volume in this test (only 100 documents), model loading time accounted for a larger proportion of the total time. In real-world scenarios, YANG models are only loaded after a restart, which will not cause a performance issue.

In DB4YANG, document parsing and data tree serialization have become performance bottlenecks. During document insertion and reading, the functions `lyd_parse_data_mem` and `lyd_print_mem` occupied 18.19% and 3.01% of the program's runtime, respectively. Together, they accounted for approximately 93% of the time spent in the document insertion and reading phase, which itself took up 22.82% of the total runtime. The function `lyd_parse_data_mem` is used to load document data of a certain format (e.g., LYB, JSON, or XML) into a data tree in memory, while `lyd_print_mem` converts the data tree into document data (e.g., JSON or XML). For simplicity, these two functions are referred to as the document parsing function and the data tree serialization function, respectively, in the following discussion.

Analysis of Fig. 10 reveals that the Sysrepo functions `sr_modinfo_data_load`, `lyd_validate_modules`, and `sr_modinfo_data_store` are primarily responsible for the increased execution time. In large-scale configuration scenarios, the volume of saved configuration data significantly increases. Consequently, this leads to longer processing time for data loading, module validation, and data storage.

6 Conclusions

This paper focuses on resource-limited settings such as base stations and explores the implementation of an embedded database system for managing YANG model data. We have designed and implemented high-speed read and write operations tailored for YANG model management and its associated data. We propose optimization methods to mitigate performance degradation caused by using relational databases for YANG model data, which involves non-native processing and format conversion. This directly addresses the technical hurdles of traditional methods, such as storing YANG model data in relational tables, which requires conversion between 2D relational data and YANG model data, incurring unnecessary overhead.

Furthermore, we present strategies for creating a lightweight, high-performance indexing system. This system not

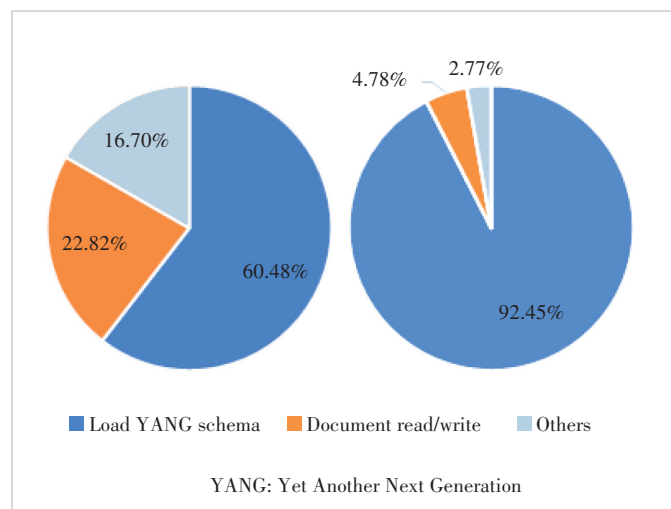


Figure 10. Perf test of DB4YANG and Sysrepo

only boosts data access speeds but also excels in space efficiency, making it ideal for handling extensive datasets. Through effective index structures, we have achieved swift data retrieval and enhanced query performance while keeping storage usage in check, making it suitable for environments with limited resources.

Additionally, this paper delves into transaction processing for YANG model data, adhering strictly to ACID principles to ensure data integrity—a critical factor in databases facing high concurrency and complex operations. Our data management module operates on a transactional basis. Data is valid upon transaction commit and void if the transaction is rolled back. We have developed a transaction management system for DB4YANG to fulfill ACID requirements.

Experimental results indicate that DB4YANG outperforms existing methods in terms of accuracy and efficiency. Moreover, its transaction management meets industry standards and demonstrates its viability for real-world applications. However, our research has certain limitations. For instance, performance is somewhat constrained by the parsing speed of YANG model documents using the `libyang` library. This can lead to bottlenecks during document parsing and conversion. In future work, we aim to enhance the database's performance and scalability by rewriting the YANG model's parsing and processing functions.

References

- [1] Yi X X, Zhang N H, Liu Y C, et al. Key technologies for intelligent computing power network [J]. ZTE technology journal, 2025, 31(2): 31–38. DOI: 10.12142/ZTETJ.202502005
- [2] Zhu L H, Karim M M, Sharif K, et al. SDN controllers: a comprehensive

¹ www.brendangregg.com/linuxperf.html

- analysis and performance evaluation study [J]. *ACM computing surveys*, 2021, 53(6): 1 – 40. DOI: 10.1145/3421764
- [3] Wang H. Improvement and implementation of wireless network topology system based on SNMP protocol for router equipment [J]. *Computer communications*, 2020, 151: 10 – 18. DOI: 10.1016/j.comcom.2019.12.038
- [4] Bhat M G, Bhattacharjee S, Gündoğan C, et al. CORECONF, NETCONF, and RESTCONF: benchmarking network orchestration in constrained IIoT devices [J]. *IEEE Internet of Things journal*, 2024, 11(7): 13082 – 13090. DOI: 10.1109/jiot.2023.3338470
- [5] Györfi C A, Dumş-Burescu D V, Zmaranda D R, et al. A comparative study of MongoDB and document-based MySQL for big data application data management [J]. *Big data and cognitive computing*, 2022, 6(2): 49. DOI: 10.3390/bdcc6020049
- [6] Golmohammadi Z. Centralized model driven trace route mechanism for TCP/IP routers: remote traceroute invocation using NETCONF API and YANG data model [D]. Tampere: Tampere University, 2019
- [7] Lee Y M, Lee J K. YANG-based data modeling techniques for the content layer of NETCONF to improve query throughput [J]. *Journal of Internet technology*, 2018, 19(7): 1993 – 2007. DOI: 10.3966/160792642018121907004
- [8] Dallaglio M, Sambo N, Cugini F, et al. Control and management of transponders with NETCONF and YANG [J]. *Journal of optical communications and networking*, 2017, 9(3): B43. DOI: 10.1364/jocn.9.000b43
- [9] Amlou A, Abane A, Merzouki M, et al. Automated network programmability using OpenConfig YANG models and NETCONF protocol [C]//The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). IEEE, 2023: 1 – 5. DOI: 10.1109/aiccsa59173.2023.10479244
- [10] Liu Y B, Dong Y Q, Shen J, et al. MAYA: exploring multi-form attributes of node to align YANG data models [J]. *Computer networks*, 2023, 222: 109538. DOI: 10.1016/j.comnet.2022.109538
- [11] Martinovic B, Vuleta M, Popic S, et al. Simple verification improvements of the NETCONF data model in programmable computer Networks [C]//The 30th Telecommunications Forum (TELFOR). IEEE, 2022: 1 – 4. DOI: 10.1109/telfor56187.2022.9983688
- [12] Guliato D, de Melo E V, Rangayyan R M, et al. POSTGRESQL-IE: an image-handling extension for PostgreSQL [J]. *Journal of digital imaging*, 2009, 22(2): 149 – 165. DOI: 10.1007/s10278-007-9097-5
- [13] Marwedel P. Embedded system design: embedded systems foundations of cyber-physical systems, and the Internet of Things [M]. Cham: Springer International Publishing, 2021. DOI: 10.1007/978-3-030-60910-8
- [14] Narayanan V, Xie Y. Reliability concerns in embedded system designs [J]. *Computer*, 2006, 39(1): 118 – 120. DOI: 10.1109/mc.2006.31
- [15] Gaffney K P, Prammer M, Brasfield L, et al. SQLite: past, present, and future [J]. *Proceedings of the VLDB endowment*, 2022, 15(12): 3535 – 3547. DOI: 10.14778/3554821.3554842
- [16] Wang L, Ding G Q, Zhao Y L, et al. Optimization of LevelDB by separating key and value [C]//The 18th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT). IEEE, 2017: 421 – 428. DOI: 10.1109/pdcat.2017.00074
- [17] Dong S Y, Kryczka A, Jin Y Q, et al. RocksDB: evolution of development priorities in a key-value store serving large-scale applications [J]. *ACM transactions on storage*, 2021, 17(4): 1 – 32. DOI: 10.1145/3483840
- [18] Chauhan A. A review on various aspects of mongodb databases [J]. *International journal of engineering research & technology (IJERT)*, 2019, 8 (5): 90 – 92
- [19] Bai Y. JDBC API and JDBC drivers [M]//Oracle database programming with Java: ideas, designs, and implementations. Boca Raton: Auerbach Publications, 2022: 67 – 86. DOI: 10.1201/9781003304029-3
- [20] Sysrepo Project. Sysrepo.org [EB/OL]. [2024-02-14]. <http://www.sysrepo.org>
- [21] Ahearne S, Verbishchuk Y, Sreenan C, et al. Software defined control of tunable optical transceivers using NETCONF and YANG [C]// European Conference on Networks and Communications (EuCNC). IEEE, 2018: 1 – 86. DOI: 10.1109/eucnc.2018.8443188

Biographies

Han Bing is currently pursuing his MS degree in electronic information at the National Elite Engineers School, Harbin Institute of Technology, China. He specializes in research directions such as database system design and edge storage.

Gao Jirun is currently pursuing his MS degree in the Department of Computing, Harbin Institute of Technology, China. He focuses on research areas such as database systems and edge storage.

Zhang Kaiqi (zhangkaiqi@hit.edu.cn) is a lecturer at the Department of Computing, Harbin Institute of Technology, China. His research interests include spatiotemporal big data query, approximate computation, and embedded database systems. He has published 15 papers.

Wang Hongzhi is a professor at the Department of Computing, Harbin Institute of Technology, China. His research interests include big data governance, big data management and analysis, database systems, and industrial big data. He has published more than 300 papers.

Chen Hedui is a senior engineer at ZTE Corporation. He specializes in databases, new hardware acceleration, lakehouse integration, and AI for databases.



A Method for Text Information Extraction Leveraging Large Language Models and Two-Stage Information Filtering

Lu Ping¹, Qin Bin², Wang Yiyang², Xu Yan³,
Deng Fangwei¹

(1. ZTE Corporation, Shenzhen 518057, China;
2. Shenzhen University, Shenzhen 518060, China;
3. Beijing Xingyun Digital Technology Co., Ltd., Beijing 100176, China)

DOI: 10.12142/ZTECOM.202603011

<https://kns.cnki.net/kcms/detail/34.1294.TN.20260826.1349.004.html>,
published online August 26, 2026

Manuscript received: 2024-09-20

Abstract: This paper addresses shortcomings in the current combined application of large language models (LLMs) and vector knowledge bases by proposing a novel method for text information extraction. The method integrates the powerful language understanding capabilities of LLMs with the robust matching ability of two-stage information filtering, aiming to improve both the hit rate for retrieving relevant background knowledge and the accuracy of text information extraction. Experimental results on a real-world accident investigation report dataset demonstrate that the proposed method achieves a significant 37% improvement in extraction accuracy compared to baseline methods. Additionally, comparative results across data from different industries further validate the method's strong generalization performance, suggesting its potential as an effective technical solution for other long-text information extraction tasks.

Keywords: large language models; two-stage information filtering; text information extraction

Citation (Format 1): Lu P, Qin B, Wang Y Y, et al. A method for text information extraction leveraging large language models and two-stage information filtering [J]. *ZTE Communications*, 2026, 24(3): 92 – 98. DOI: 10.12142/ZTECOM.202603011

Citation (Format 2): P. Lu, B. Qin, Y. Y. Wang, et al., “A method for text information extraction leveraging large language models and two-stage information filtering,” *ZTE Communications*, vol. 24, no. 3, pp. 92 – 98, Sept. 2026. doi: 10.12142/ZTECOM.202603011.

1 Introduction

Text information extraction is the process of identifying and extracting key information from textual data based on predefined labels. In the current era of information overload, immense value is hidden within vast amounts of data. Efficiently extracting useful information from this data has become a key research focus in fields such as computer science, data mining, and information retrieval. However, in the context of big data, traditional methods of text information extraction face significant challenges, including high redundancy and low accuracy. Thus, developing more efficient and precise extraction technologies is crucial.

In recent years, the rapid advancement of artificial intelligence technology, spurred by the surge of large language models (LLMs) such as ChatGPT, has gained global traction. Approaches based on LLMs are increasingly becoming a new trend in text information extraction research. These large models typically consist of hundreds of millions to tens of bil-

lions of parameters, offering significantly greater learning and representational capabilities compared to traditional smaller models. After being trained on extensive datasets, they can more effectively capture complex patterns and features within text data, thus enhancing the accuracy of information extraction. These methods often frame predefined labels as questions, organizing the questions and raw text data in a specific format before inputting them into the model. The extraction results are then obtained by querying the model in this structured manner.

However, due to the current limitations of LLMs in processing long texts, even models that claim to handle long texts effectively can still experience a significant drop in accuracy when dealing with ultra-long textual data in real-world vertical domain tasks, and may even exhibit phenomena such as model hallucinations. To address this issue in the task of text information extraction, researchers have proposed a solution that integrates LLMs with vector knowledge bases. The core idea of this approach is to leverage the powerful language understanding capabilities of LLMs in combination with the precise information provided by the vector knowledge base, thereby achieving complementary advantages. The approach

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. IA20231030016.

starts with vector retrieval, which extracts semantically relevant background knowledge from the knowledge base that closely aligns with the question. This background knowledge is then combined with the question and fed into the LLM. By focusing only on a limited amount of original text that is highly relevant to the question, this approach aims to improve the reliability of the model's responses.

Nevertheless, background knowledge retrieval based on semantic similarity relies on the proximity between the question and the background knowledge in semantic space. This proximity does not always ensure content relevance, leading to a low hit rate, which has become a major limitation of this method. Additionally, it is often difficult to strike a balance between the quantity and quality of candidate entries generated through a single filtering step, which ultimately impacts the overall system performance.

To overcome these limitations, this paper proposes a novel approach—text information extraction based on LLMs with two-stage information filtering. This method decomposes the question of predefined label construction into multiple retrieval terms and an inquiry term. Each retrieval term is used exclusively for the initial retrieval, and a set operation is performed on all the initial results to ensure the uniqueness of text blocks. Finally, the background knowledge obtained through the two-stage filtering process is fed into the LLM, producing high-quality outputs. The advantages of this method are highlighted in four key aspects:

- **Higher background knowledge retrieval hit rate:** By designing multiple retrieval terms for the initial retrieval, the results are more targeted and better aligned with the practical needs of the information extraction task. Since different retrieval terms focus on different aspects of the question, this method covers potentially relevant background knowledge more comprehensively, thereby improving the hit rate.
- **Increased accuracy of text information extraction:** The inquiry term is used to further filter the background knowledge matched during the initial retrieval, selecting information that is more highly correlated with the question. This process ensures that the final retrieved background knowledge is not only semantically related to the question but also more contextually aligned with the actual needs of the text information extraction task, thereby enhancing accuracy.
- **Reduced model hallucinations:** By applying two-stage filtering and re-ranking mechanisms to further screen the background knowledge, this approach ensures that the final background information passed to the LLM is highly relevant to the information extraction task, helping to reduce model hallucinations.
- **Improved system efficiency:** The two-stage filtering and re-ranking mechanisms further streamline the background knowledge that is passed to the LLM, reducing the amount of information that the model needs to process, thus lowering computational burdens and improving overall system efficiency.

2 Related Work

2.1 Evolution of LLMs

Since the introduction of deep learning algorithms in 2006, they have achieved remarkable success in image recognition and found widespread application in natural language processing (NLP). During this critical period, models such as Word2Vec^[1], Seq2Seq^[2], and the recurrent neural network (RNN) Encoder-Decoder^[3] laid the theoretical and practical foundations for the subsequent development of LLMs.

In 2017, Google's research team introduced the Transformer^[4], whose efficient parallel processing and ability to handle long-range dependencies made it feasible to build LLMs. Based on the Transformer, Devlin et al. proposed BERT^[5], which significantly improved contextual understanding through bidirectional pre-training. BERT and its derivatives, such as RoBERTa^[6] and ALBERT^[7], achieved breakthrough results in NLP tasks. BERT-BiLSTM-ATT^[8] was specifically optimized for Chinese, enhancing its performance in understanding and generating Chinese text. Pre-trained models have now become the new paradigm in NLP, ushering in the era of large-scale pre-trained models in the field of artificial intelligence^[9].

Since 2020, large-scale pre-trained models like GPT-3^[10] have demonstrated exceptional capabilities in areas such as natural language generation, conversational systems, and text summarization, thanks to their massive parameter scale and outstanding performance. The widespread application of these models has not only gained attention in academia but also shown their practicality in industries such as finance, healthcare, and law, significantly improving efficiency and advancing intelligent services. As a result, LLMs have officially entered the stage of production applications, becoming a crucial tool for enhancing productivity and marking the transition from academic research to industrial applications.

2.2 Application and Development Frameworks of LLMs

With the continuous advancement of LLMs, their applications have extensively permeated various sectors of society, offering transformative solutions. For instance, the intelligent legal system Disc-Law^[11] leverages LLMs to efficiently process legal documents and provide consultation services, significantly enhancing the efficiency and accessibility of legal services. ChatGPT, known for generating fluent and logically coherent language, has been widely used in fields such as content creation, data analysis, and content refinement. The applications of Chinese LLMs, such as Baidu Wenxin Yiyan, Zhipu Qingyan, and iFlytek SPARK, mark a new milestone in Chinese language processing capabilities.

To simplify the development and deployment of LLMs, specialized application development frameworks such as Semantic Kernel and LangChain have emerged. These frameworks provide a suite of tools and interfaces that help developers eas-

ily integrate and apply LLMs. Semantic Kernel focuses on deeply customizing semantic understanding capabilities, while LangChain streamlines the construction of complex language processing workflows through modular design. Leveraging these frameworks to integrate LLMs with vector knowledge bases lays a solid foundation for applying LLMs to domains such as emergency management.

2.3 Text Information Extraction

In the field of NLP, Named Entity Recognition (NER)^[12] is a sub-task of text information extraction, specifically aimed at identifying entities in text that have specific meanings, such as names of people and locations. Through well-established NER methods, we can extract important information from text, highlight key points, and filter out redundant background information. However, real-world text information extraction demands are often complex and diverse. Using NER for information extraction can be limited by predefined text types and entity categories. For example, in scenarios such as extracting information from emergency plans, the diversity of required information may exceed the predefined entity categories of a model. Moreover, fine-tuning models for a large amount of information can consume substantial time and resources, creating bottlenecks in practical applications. As a result, traditional NER methods often struggle to handle challenging tasks like extracting information from emergency plans.

In recent years, the rapid development of Retrieval-Augmented Generation (RAG) technology has provided a novel approach to tackling complex information extraction tasks. By incorporating relevant information retrieved from external knowledge bases into LLMs, RAG significantly enhances the accuracy of generated content while reducing the risk of producing misleading information. Building on the traditional RAG framework, several advanced methods have

been proposed to address the challenges of processing long texts. For instance, MemoRAG^[13] introduces a memory-inspired knowledge discovery mechanism, demonstrating exceptional performance in handling long texts. LightRAG^[14] integrates knowledge graphs, leveraging graph structures to represent entities and their relationships, thereby enhancing contextual awareness and mitigating the issue of fragmented answers caused by reliance on flat data structures. ChunkRAG^[15], on the other hand, segments long texts into semantically coherent chunks and utilizes LLMs to filter out irrelevant chunks prior to the generation phase. This approach minimizes the interference of irrelevant information, thereby improving the accuracy of generated responses.

Inspired by these advancements, we applied a collaborative strategy combining LLMs and vector-based knowledge bases to information extraction tasks. This approach leverages the language model's strengths in understanding while integrating precise information provided by the vector knowledge base, achieving more efficient and accurate information extraction. Additionally, drawing on the divide-and-conquer strategy underlying ChunkRAG, our method breaks down complex queries and filters content in stages, effectively addressing the limitations of LLMs in processing long texts.

3 Proposed Method

3.1 Overview

The overall architecture of the proposed method is illustrated in Fig. 1. Specifically, this approach takes the document and the information labels to be extracted as inputs. First, a vector knowledge base is constructed based on the document. Then, an LLM is employed to generate a set of retrieval terms and inquiry terms corresponding to the information labels. Subsequently, the vectorized retrieval term set is

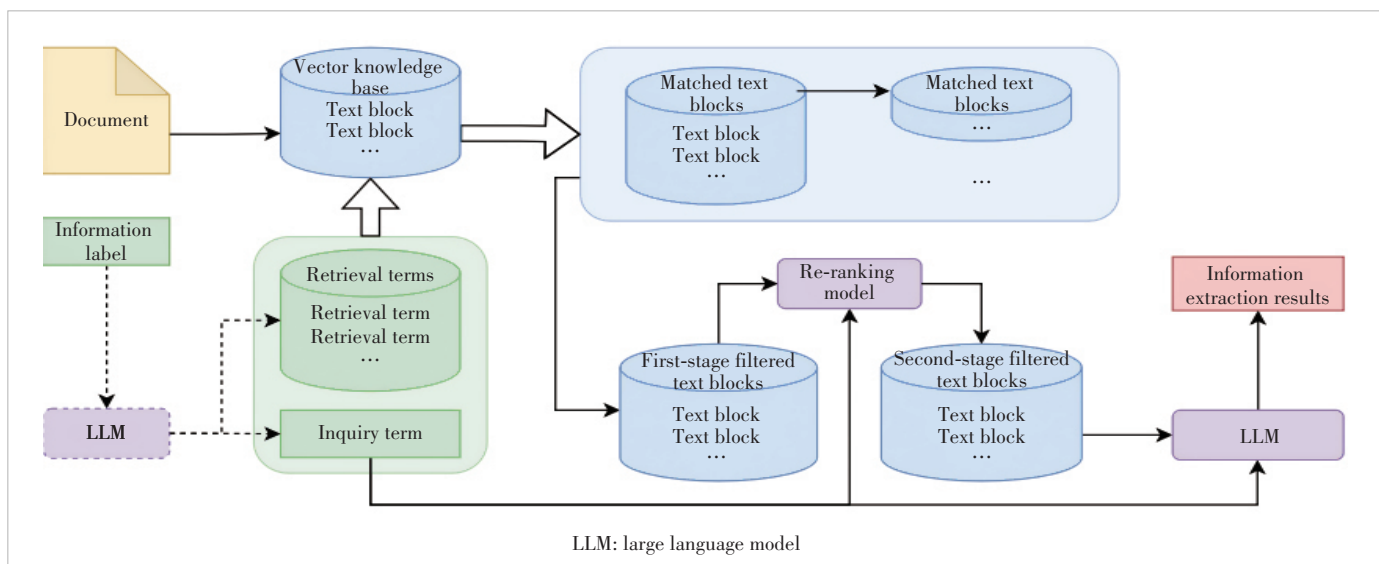


Figure 1. Framework of text information extraction method based on LLMs and two-stage information filtering

used sequentially to perform text block matching within the vector knowledge base. The resulting set of matched text blocks undergoes two-stage information filtering through semantic relevance matching with the inquiry terms. Finally, the filtered background knowledge and inquiry terms are input into the LLM, which generates the text information extraction results. The following section provides a detailed description of this process.

3.2 Vector Knowledge Base Construction

1) Long text segmentation

When processing long texts, we adopt a recursive character-based text splitting method to ensure that the length of the resulting text blocks does not exceed a predefined character limit. First, the given text is split into several text blocks at natural boundaries such as sentences or paragraphs. This initial splitting maximizes the preservation of the text's semantic integrity. Then, for each text block, if its length exceeds the character limit, the block is further recursively split until all text blocks satisfy the character limit requirement. The recursive process can be represented as the function $\text{split}(T, L)$, which accepts the text and character limit as inputs and returns a set of sub-blocks that meet the character limit:

$$\text{split}(T, L) = \begin{cases} \{T\}, & |T| \leq L \\ \bigcup_{i=1}^n \text{split}(T_i, L), & \text{otherwise} \end{cases} \quad (1)$$

where T represents the input text, L represents the character limit, and T_i represents the i -th text block obtained by splitting T .

2) Text block vectorization

At the vectorization stage, we use a pre-trained embedding model to generate high-dimensional vector representations of the text blocks. For the set of text blocks $\{T_i | T_i \in \text{split}(T, L)\}$, returned from splitting the long text, each text block T_i is converted into a vector representation.

$$\mathbf{v}_i = f(T_i), \forall T_i \in \text{split}(T, L) \quad (2)$$

where the embedding function $f(\cdot)$ is implemented by the pre-trained embedding model, which directly maps text

blocks to high-dimensional vector representations $\mathbf{v}_i$. This vectorization method preserves the semantic information of the text blocks, facilitating subsequent retrieval operations in the semantic space.

3) Knowledge base construction

When constructing the vector knowledge base, we first store the vector set $V = \{\mathbf{v}_i | T_i \in \text{split}(T, L)\}$ in the vector knowledge base M , and then build an index for each vector. This process can be expressed as:

$$M = \left(\text{store}_{\text{vectors}}(V), \left\{ \text{index}_{\text{vector}}(\mathbf{v}_i) \mid \mathbf{v}_i \in V \right\}, d(\cdot) \right) \quad (3),$$

where $\text{store}_{\text{vectors}}(V)$ denotes storing the vector set V in the database, $\{\text{index}_{\text{vector}}(\mathbf{v}_i) \mid \mathbf{v}_i \in V\}$ denotes indexing each vector, and $d(\cdot)$ represents the similarity measurement function of M . In our implementation, we use the Euclidean distance as the similarity measurement, which is defined as:

$$d(\mathbf{v}_i, \mathbf{v}_j) = \sqrt{\sum_{k=1}^m (\mathbf{v}_i[k] - \mathbf{v}_j[k])^2} \quad (4),$$

where $\mathbf{v}_i[k]$ denotes the k -th element of the text vector $\mathbf{v}_i$.

3.3 Two-Stage Information Filtering

We propose a novel background knowledge retrieval mechanism that decomposes the question into multiple retrieval terms and a separate inquiry term. By utilizing initial retrieval followed by two-stage information filtering, we enhance the retrieval hit rate of background knowledge.

3.3.1 Initial Retrieval

First, we extract multiple retrieval terms q_{search_k} from the question, with each retrieval term focusing on different aspects of the question to cover a broader range of potential background knowledge, thereby enhancing the initial search's coverage. In this method, the separation of retrieval terms and the inquiry term is aided by LLMs, with some manual intervention. For the same information label, the retrieval terms obtained from the initial separation can be embedded with the inquiry term for direct use in subsequent steps. Table 1 presents

Table 1. An example of the separation of retrieval and inquiry terms

Information Label	Before Decoupling	After Decoupling	
		Retrieval Term	Inquiry Term
Accident location	In which province, city, county (district), and street is the accident site located, and what is the exact address?	1) In which province, city, county (district), and street is the accident site located, and what is the exact address?	2) Accident location
		3) Place of the accident	4) Position of the accident
		5) Accident site	6) Area of the accident
		7) Detailed address of the accident	8) Specific location of the accident
		9) Address of the accident	...
			What is the specific location of the accident?

an example of this separation process.

Next, each retrieval term is converted into a high-dimensional vector representation $\mathbf{v}_{q_{\text{search}_i}}$:

$$\mathbf{v}_{q_{\text{search}_i}} = f(q_{\text{search}_i}) \quad (5)$$

During the similarity search process, we filter the top N results whose vector distances to each retrieval term are below the threshold S . The algorithmic formula is as follows:

$$s = \left\{ \mathbf{v}_i \mid d(\mathbf{v}_{q_{\text{search}_i}}, \mathbf{v}_i) \leq S \right\} \cap \left\{ \mathbf{v}_i \mid \mathbf{v}_i \in M \leq N \right\} \quad (6)$$

where $\mathbf{v}_i$ represents the vector representation of the text blocks stored in the vector knowledge base M .

Utilizing the similarity search algorithm s , we obtain the set of retrieval results in a vector format R_k :

$$R_k = s(\mathbf{v}_{q_{\text{search}_k}}, N, S) \quad (7)$$

We then convert the retrieved vector set R_k into the corresponding set of text blocks T_{R_k} :

$$T_{R_k} = \{T_i \mid \mathbf{v}_i \in R_k\} \quad (8)$$

Finally, we merge all text block sets T_{R_k} to ensure the uniqueness of each text block, forming a filtered set of text blocks T_R :

$$T_R = \bigcup_k T_{R_k} \quad (9)$$

Through the background knowledge retrieval mechanism that separates retrieval terms and inquiry terms, we ensure that the search results meet both the quantitative constraints and the similarity score requirements, thereby enhancing the accuracy and relevance of the retrieval.

3.3.2 Second-Stage Information Filtering

In the initial retrieval stage, we obtain a set of first-stage filtered text blocks T_R related to the retrieval term q_{search_i} . Although these results have undergone initial filtering, they may still contain a significant number of entries with low relevance. Directly passing this background knowledge in its entirety to the LLM is inappropriate, as it not only increases computational overhead but may also expose the model to noisy information, resulting in hallucinations. To further enhance the relevance of the retrieved background information, we introduce a second-stage information filtering process.

In the second-stage filtering phase, we use a pre-trained re-ranking model to further refine and rank the initially filtered background knowledge.

First, we compute the relevance score of each text block obtained from the initial filtering with respect to the inquiry term:

$$\text{score}(q_{\text{ask}}, T_i) = \text{rerank}(q_{\text{ask}}, T_i) \quad (10)$$

where q_{ask} denotes the inquiry term and T_i represents the i -th text block in the set T_R . The function $\text{rerank}(\cdot)$ is implemented using a pre-trained re-ranking model.

Next, we sort the elements of the text block set T_R based on their relevance scores and select the top K text blocks with scores above the threshold S' as the final background knowledge matching result T_{final} :

$$T_{\text{final}} = T_i \mid \text{score}(q_{\text{ask}}, T_i) \geq S' \cap T_i \mid |T_{\text{final}}| \leq K \quad (11)$$

The second-stage filtering further removes low-relevance entries from the initial filtering results, significantly improving the matching degree between the background knowledge and the inquiry term q_{ask} . This process helps us precisely target critical background knowledge, avoiding the transmission of a large volume of irrelevant information to the LLM, thus ensuring the quality and efficiency of the model's responses.

3.4 LLM Interaction

After obtaining the high-relevance two-stage filtered text block set T_{final} , we format it with the inquiry term q_{ask} using a prompt template and feed this into the LLM for final answer generation. This process can be represented as follows:

$$A = \text{LLM}(\text{template}(q_{\text{ask}}, T_{\text{final}})) \quad (12)$$

where the function $\text{template}(\cdot)$ refers to the operation of applying the prompt template, and the function $\text{LLM}(\cdot)$ is implemented by the LLM. Specifically, this method draws on the prompt engineering guidelines published by OpenAI^[16]. Based on empirical principles such as using delimiters to distinguish instructions from context and clarifying the desired output format through examples, a prompt template as illustrated in Fig. 2 has been designed. This prompt template aims to assist the LLM in better understanding the requirements of the information extraction task and generating extraction results that conform to the expected format.

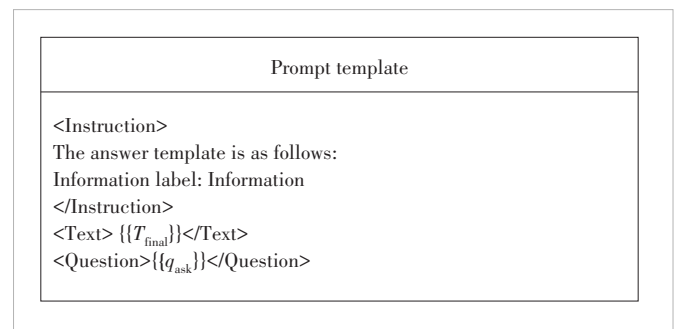


Figure 2. Prompt template for LLM

4 Experiment

4.1 Dataset

The dataset used in this study consists of 219 authentic Chinese accident investigation reports, encompassing two major industry categories: construction and commercial manufacturing, as well as several subcategories. The industry distribution is shown in Table 2.

Considering the focus on accident information in the field of emergency management, as well as the common information found in accident investigation reports, this study aims to extract four types of information from the reports, as shown in Table 3. In total, 876 information items need to be extracted from the 219 accident investigation reports.

4.2 Implementation Details

We developed an information extraction application for accident investigation reports based on the proposed method, using the LangChain and Streamlit frameworks. Fig. 3 illustrates the implementation process of the application. The selected LLM is ChatGLM3-6B^[17], which supports both Chinese and English with an 8k context length. The temperature parameter is set to

Table 2. Industry distribution of the dataset

Major Industry Category	Subdivision	Number of Reports
Construction	Residential building	48
	Civil engineering	31
Commercial manufacturing	Trade and industry	100
	Chemical industry	40

Table 3. Information labels for accident investigation reports

Index	Information Label
1	Accident location
2	Casualties
3	Economic loss
4	Legal basis

0.01. A lower temperature value means the model is more inclined to choose high-probability words during sampling, thereby generating more deterministic and consistent outputs. This aligns with the high accuracy and consistency requirements of the information extraction task. The embedding model is M3E-base^[18], the re-ranking model is BGE-reranker-v2-m3^[19], and the vector knowledge base is built using Faiss. This application is deployed on a single RTX-4090 GPU. The settings for the various parameters in the method are listed in Table 4.

4.3 Comparative Experiment

To validate the superiority of the proposed method, we conducted comparative experiments using Langchain-Chatchat^[20], a representative baseline method within the traditional LLM and vector knowledge base collaborative application framework. The blue section in Fig. 3 highlights the improvements of our method compared to the baseline. Table 5 presents the results of the comparative experiments, using accuracy as the evaluation metric.

As shown in Table 5, when extracting information from real-world Chinese accident investigation reports, the proposed method significantly outperformed Langchain-Chatchat in the

Table 4. Parameter configuration in the method

Parameter	Value
Similarity threshold S	0.5
First-stage filtering text block threshold N	20
Semantic similarity threshold S'	0.5
Second-stage filtering text block threshold K	8

Table 5. Comparative experiment results

Method	Accident Location	Casualties	Economic Loss	Legal Basis	Average
Langchain-Chatchat	0.47	0.31	0.52	0.13	0.36
Ours	0.76	0.75	0.67	0.75	0.73

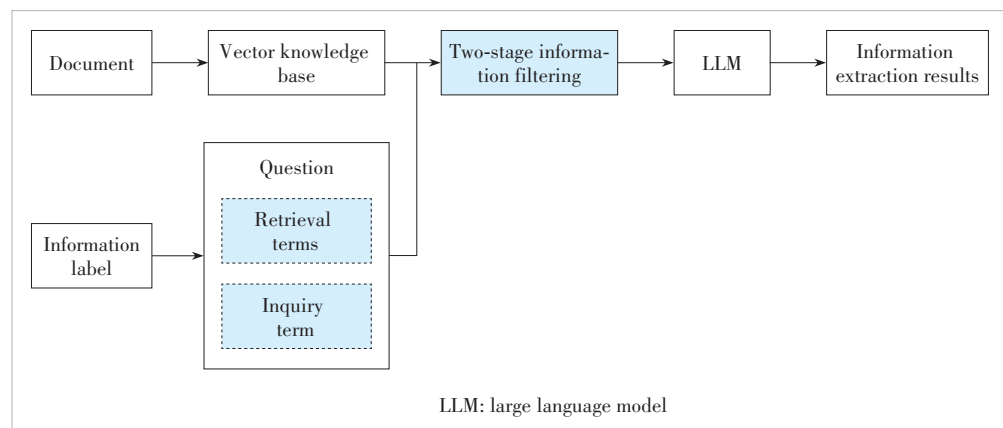


Figure 3. Flowchart of information extraction process from accident investigation reports

four information categories: location of the accident, casualty situation, economic loss, and legal basis. Notably, the accuracy for extracting legal basis improved from 13% to 75%, and the average extraction accuracy for the four categories increased from 36% to 73%.

Table 6 illustrates the differences in extraction accuracy between the two methods for accident investigation reports from various industries. This

Table 6. Experimental results across different industries

Method	Residential Building	Civil Engineering	Trade and Industry	Chemical Industry	Average
Langchain-Chatchat	0.41	0.27	0.40	0.50	0.40
Ours	0.73	0.75	0.79	0.62	0.72

indicates that the proposed method demonstrates greater generalization across different sectors.

5 Conclusions

This paper proposes a text information extraction method based on LLMs and two-stage information filtering. Validation on a real-world accident investigation report dataset demonstrates that this method significantly improves the accuracy of information extraction compared to the baseline method, while also mitigating model hallucinations. Moreover, comparative results across documents from different industries further confirm that the proposed method possesses considerable generalization capability, rather than serving merely as an optimized solution tailored to specific text types. This provides a technical reference for the practical application of long-text information extraction in other domains.

References

- [1] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [PP/OL]. arXiv [2024-09-20]. <https://arxiv.org/pdf/1301.3781v1>
- [2] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]//The 28th International Conference on Neural Information Processing Systems. ACM, 2014: 3104 – 3112. DOI: 10.5555/2969033.2969173
- [3] Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]//The 2014 Conference on Empirical Methods in Natural Language Processing. ACL, 2014: 1724 – 1734. DOI: 10.3115/v1/d14-1179
- [4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//The 31st International Conference on Neural Information Processing System. ACM, 2017: 6000 – 6010. DOI: 10.5555/3295222.3295349
- [5] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2019: 4171 – 4186. DOI: 10.18653/v1/N19-1423
- [6] Liu Y H, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pre-training approach [PP/OL]. arXiv (2019-07-26) [2024-09-20]. <https://arxiv.org/abs/1907.11692>
- [7] Lan Z Z, Chen M D, Goodman S, et al. ALBERT: a lite BERT for self-supervised learning of language representations [PP/OL]. V6. arXiv (2020-02-09) [2024-09-20]. <https://arxiv.org/abs/1909.11942>
- [8] Li D Y, Tu Y F, Zhou X S, et al. End-to-end Chinese entity recognition based on BERT-BiLSTM-ATT-CRF [J]. ZTE communications, 2022, 20 (S1): 27 – 35. DOI: 10.12142/ZTECOM.2022S1005
- [9] Zhou C, Li Q, Li C, et al. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT [PP/OL]. V4. arXiv (2023-05-01) [2024-09-20]. <https://arxiv.org/abs/2302.09419>
- [10] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners [PP/OL]. V4. arXiv (2020-07-22) [2024-09-20]. <https://arxiv.org/abs/2005.14165>
- [11] Yue S B, Chen W, Wang S Y, et al. DISC-LawLLM: fine-tuning large language models for intelligent legal services [PP/OL]. V2. arXiv (2023-09-23) [2024-09-20]. <https://arxiv.org/abs/2309.11325>
- [12] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs [J]. Transactions of the association for computational linguistics, 2016, 4: 357 – 370. DOI: 10.1162/tacL_a_00104
- [13] Qian H J, Liu Z, Zhang P T, et al. MemoRAG: boosting long context processing with global memory-enhanced retrieval augmentation. [PP/OL]. arXiv (2024-09-09) [2024-09-20]. <https://arxiv.org/abs/2409.05591>
- [14] Guo Z R, Xia L H, Yu Y H, et al. LightRAG: simple and fast retrieval-augmented generation [C]//Proceedings of Findings of the Association for Computational Linguistics. ACL, 2025: 10746 – 10761. DOI: 10.18653/v1/2025.findings-emnlp.568
- [15] Singh I S, Aggarwal R, Allahverdiyev I, et al. ChunkRAG: novel LLM-chunk filtering method for RAG systems [EB/OL]. V1. arXiv (2024-10-25) [2024-10-30]. <https://arxiv.org/abs/2410.19572>
- [16] OpenAI. Best practices for prompt engineering with the OpenAI API [EB/OL]. [2024-09-20]. <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api>
- [17] Du Z X, Qian Y J, Liu X, et al. GLM: general language model pretraining with autoregressive blank infilling [PP/OL]. V2. arXiv (2022-03-17) [2024-09-20]. <https://arxiv.org/abs/2103.10360>
- [18] Wang Y X, Sun Q X, He S C. M3e-base [EB/OL]. (2023-06-07) [2024-09-20]. <https://huggingface.co/moka-ai/m3e-base>
- [19] Gao L, Callan J. Condenser: a pre-training architecture for dense retrieval [PP/OL]. V2. arXiv (2021-09-20) [2024-09-20]. <https://arxiv.org/abs/2104.08253>
- [20] Chatchat-Space. Langchain-Chatchat. [EB/OL]. [2024-09-20]. <https://github.com/chatchat-space/Langchain-Chatchat>

Biographies

Lu Ping is the vice president of ZTE Corporation. He also serves as the deputy director of the State Key Laboratory of Mobile Network and Mobile Multimedia Communication. His research interests include artificial intelligence, XR, immersive communication, and network-media convergence.

Qin Bin (qinbin@szu.edu.cn) is a professor-level senior engineer of Shenzhen University, China. His research interests focus on artificial intelligence.

Wang Yiyang received her master's degree in new-generation information technology from Shenzhen University, China in 2026. Her research interests include the applications of large language models and 3D vision.

Xu Yan holds an MS degree in mechanical engineering. He is currently a product manager at Beijing Xingyun Numerical Technology Co., Ltd. His research interest focuses on the digital application of AI technology in the industry.

Deng Fangwei is a chief research engineer of ZTE Corporation. Her research interests include artificial intelligence, XR, immersive communication, and network-media convergence.

The 2nd Youth Expert Committee

for Promoting Industry-University-Institute Cooperation

Director **Chen Wei**, Beijing Jiaotong University

Deputy Director **Qin Xiaoqi**, Beijing University of Posts and Telecommunications

Lu Dan, ZTE Corporation

Members (Surname in Alphabetical Order)

Cao Jin	Xidian University
Chen Li	University of Science and Technology of China
Chen Qimei	Wuhan University
Chen Shuyi	Harbin Institute of Technology
Chen Siheng	Shanghai Jiao Tong University
Chen Wei	Beijing Jiaotong University
Gao Zhen	Beijing Institute of Technology
Guan Ke	Beijing Jiaotong University
Han Chong	Shanghai Jiao Tong University
Han Kaifeng	China Academy of Information and Communications Technology
He Zi	Nanjing University of Science and Technology
Hou Tianwei	Beijing University of Posts and Telecommunications
Hu Jie	University of Electronic Science and Technology of China
Huang Chen	Purple Mountain Laboratories
Huo Jiahao	University of Science and Technology Beijing
Li Ang	Xi'an Jiaotong University
Li Li	University of Science and Technology of China
Liu Fan	Southeast University
Liu Junyu	Xidian University
Lu Dan	ZTE Corporation
Lu Youyou	Tsinghua University
Mei Weidong	University of Electronic Science and Technology of China
Ning Zhaolong	Chongqing University of Posts and Telecommunications
Pan Cunhua	Southeast University
Qi Liang	Shanghai Jiao Tong University
Qin Xiaoqi	Beijing University of Posts and Telecommunications
Qin Zhijin	Tsinghua University
Shi Yao	Harbin Institute of Technology, Shenzhen
Shi Yinghuan	Nanjing University
Tang Wankai	Southeast University
Wang Jingjing	Beihang University
Wang Xinggang	Huazhong University of Science and Technology
Wang Yongqiang	Tianjin University
Wen Miaowen	South China University of Technology
Wu Qingqing	Shanghai Jiao Tong University
Wu Yongpeng	Shanghai Jiao Tong University
Xia Wenchao	Nanjing University of Posts and Telecommunications
Xiang Luping	Nanjing University
Xu Mengwei	Beijing University of Posts and Telecommunications
Xu Tianheng	Shanghai Advanced Research Institute, Chinese Academy of Sciences
Yang Chuanchuan	Peking University
Ye Yinghui	Xi'an University of Posts and Telecommunications
Yin Haifan	Huazhong University of Science and Technology
You Changsheng	Southern University of Science and Technology
Yu Jihong	Beijing Institute of Technology
Zhang Jiao	Beijing University of Posts and Telecommunications
Zhang Jiayi	Beijing Jiaotong University
Zhang Yuchao	Beijing University of Posts and Telecommunications
Zhao Yizhe	University of Electronic Science and Technology of China
Zhao Yuda	Zhejiang University
Zhao Zhongyuan	Beijing University of Posts and Telecommunications
Zhou Yi	Southwest Jiaotong University
Zhu Bingcheng	Southeast University
Zhu Guangxu	Shenzhen Research Institute of Big Data
Zhu Zhengyu	Zhengzhou University

ZTE COMMUNICATIONS

中兴通讯技术(英文版)

ZTE Communications has been indexed in the following databases:

- Abstract Journal
- China Science and Technology Journal Database
- Chinese Journal Fulltext Databases
- Index Copernicus
- Scopus
- Ulrich's Periodicals Directory
- Wanfang Data
- WJCI 2021-2025

Industry Consultants:

Duan Xiangyang, Gao Yin, Hu Liujun, Hua Xinhai, Liu Xinyang,
Shi Weiqiang, Tu Yaofeng, Wang Huitao, Xiong Xiankui, Xu Jin,
Yan Xincheng, Zhao Yajun, Zhu Xiaoguang

ZTE COMMUNICATIONS

Vol. 24 No. 3 (Issue 96)

Quarterly

First Issue Published in 2003

Supervised by:

Anhui Publishing Group

Sponsored by:

Time Publishing and Media Co., Ltd.

Shenzhen Guangyu Aerospace Industry Co., Ltd.

Published by:

Anhui Science & Technology Publishing House

Edited and Circulated (Home and Abroad) by:
Magazine House of ZTE Communications

Staff Members:

General Editor: Wang Xiyu

Editor-in-Chief: Tao Shanyong

Executive Editor-in-Chief: Wu Ling, Huang Xinming

Deputy Editor-in-Chief: Lu Dan

Editorial Director: Wang Pingping

Editor-in-Charge: Wang Zhen, Zhu Li

Editors: Ren Xixi, Xu Ye, Yang Guangxi

Producer: Xu Ying

Circulation Executive: Wang Pingping

Assistant: Wang Kun

Editorial Correspondence:

Add: 12F Kaixuan Building, 329 Jinzhai Road,
Hefei 230061, P. R. China

Tel: +86-551-65533356

Email: magazine@zte.com.cn

Website: <http://zte.magtechjournal.com>

Annual Subscription: RMB 120

Printed by:

Anhui Tianjin Printing Technology Co., Ltd.

Publication Date: September 25, 2026

China Standard Serial Number: ISSN 1673-5188
CN 34-1294/TN