Empowering Grounding DINO with MoE: An End-to-End Framework for Cross-Domain Few-Shot Object Detection

DONG Xiugang, ZHANG Kaijin, NONG Qingpeng, JU Minhan, TU Yaofeng
(Nanjing R&D Center, ZTE Corporation, Nanjing 210012, China)

DOI: 10.12142/ZTECOM.202504009
https://kns.cnki.net/kcms/detail/34.1294.TN.20250915.1501.002.html, published online September 15, 2025
Manuscript received: 2025-08-15

Abstract: Open-set object detectors, as exemplified by Grounding DINO, have attracted significant attention due to their remarkable performance on in-domain datasets like Common Objects in Context (COCO) after only few-shot fine-tuning. However, their generalization capabilities in cross-domain scenarios remain substantially inferior to their in-domain few-shot performance. Prior work on fine-tuning Grounding DINO for cross-domain few-shot object detection has primarily focused on data augmentation, leaving broader systemic optimizations unexplored. To bridge this gap, we propose a comprehensive end-to-end fine-tuning framework specifically designed to optimize Grounding DINO for cross-domain few-shot scenarios. In addition, we propose Mixture-of-Experts (MoE)-Grounding DINO, a novel architecture that integrates the MoE architecture to enhance adaptability in cross-domain settings. Our approach demonstrates a significant 15.4 Mean Average Precision (mAP) improvement over the Grounding DINO baseline on the Roboflow20-VL benchmark, establishing a new state of the art for cross-domain few-shot object detection (CD-FSOD). The source code and models will be made available upon publication.

Keywords: cross-domain few-shot object detection; Grounding DINO; Mixture-of-Experts; open-set object detection; pseudo-labeling

Citation (Format 1): DONG X G, ZHANG K J, NONG Q P, et al. Empowering Grounding DINO with MoE: an end-to-end framework for cross-domain few-shot object detection [J]. ZTE Communications, 2025, 23(4): 77 - 85. DOI: 10.12142/ZTECOM.202504009
Citation (Format 2): X. G. Dong, K. J. Zhang, Q. P. Nong, et al., "Empowering Grounding DINO with MoE: an end-to-end framework for cross-domain few-shot object detection," *ZTE Communications*, vol. 23, no. 4, pp. 77 - 85, Sept. 2025. doi: 10.12142/ZTECOM.202504009.

1 Introduction

Open-set object detectors, such as Grounding DINO[1] and Grounding DINO 1.5[2], have attracted significant attention due to their ability to detect arbitrary objects described by natural language, even those unseen during pre-training. Leveraging vast internet-scale pre-training data, these models exhibit compelling performance on numerous in-domain datasets, such as Common Objects in Context (COCO)[3], often requiring only minimal few-shot fine-tuning to adapt to specific tasks. This capability makes them highly promising for real-world applications where exhaustive annotation is impractical.

Despite their impressive performance in in-domain scenarios, the generalization capabilities of these detectors in cross-domain settings remain a substantial challenge[4 - 5]. Cross-domain few-shot object detection (CD-FSOD) aims to recognize and localize novel objects in target domains that exhibit significant discrepancies from the source domain (i. e., the internet-scale pre-training data), using only a limited number of annotated examples. The inherent domain shift, coupled with the severe data scarcity, leads to considerable and often prohibitive performance degradation for models like Grounding DINO. This substantial drop in accuracy and reliability critically undermines their practical utility in real-world applications where target data diverges from pre-training data and the available annotations are limited, such as autonomous systems, medical diagnosis, or remote sensing. Prior work[4, 6] on adapting Grounding DINO for CD-FSOD has mainly focused on isolated techniques such as data augmentation, leaving broader systemic optimizations largely unexplored.

To address this critical gap, our work introduces the first comprehensive end-to-end fine-tuning framework specifically designed to optimize Grounding DINO for CD-FSOD. This novel framework covers a holistic pipeline, spanning from dataset construction optimization to multi-stage model training, ensuring maximal data utilization and robust adaptation.

A central and pioneering contribution of this framework is the Mixture-of-Experts (MoE)-Grounding DINO, a novel ar-

chitecture that integrates the MoE[7] module. While MoE has been widely adopted in large language models (LLMs) such as Mistral[8] and DeepSeek-V2[9], its application to open-set object detection, particularly for CD-FSOD, remains unexplored. MoE allows a model to selectively activate different "expert" sub-networks based on the input, effectively increasing the model capacity[10]. We argue that this architecture is particularly well-suited for CD-FSOD since it provides a flexible mechanism to learn more diverse representations, allowing different experts to specialize in varied visual characteristics from the limited cross-domain training data.

Experimental results demonstrate the efficacy of our proposed framework. Notably, our approach achieves a significant 15.4 Mean Average Precision (mAP) improvement over the standard Grounding DINO baseline on the challenging Roboflow20-VL[11] benchmark, establishing a new state-of-the-art model for CD-FSOD. Our contributions can be summarized as follows:

1) We propose the first end-to-end framework specifically designed for the task of CD-FSOD;

2) We are the first to integrate the MoE architecture into the domain of open-set object detection;

3) Our framework establishes a new state-of-the-art model on the Roboflow20-VL benchmark by fine-tuning Grounding DINO for CD-FSOD.

2 Related Work

2.1 Open-Set Object Detection

Compared to traditional object detection models that can only detect objects from a pre-defined set of categories, open-set object detection detects objects based on arbitrary user-provided textual category descriptions. Among influential approaches, Grounded Language-Image Pre-Training (GLIP)[12] demonstrates particularly effective performance by formulating object detection as a visual grounding task, aligning region embeddings with text embeddings. RegionCLIP[13] enhances open-set object detection by first generating region proposals via a Region Proposal Network (RPN) and then comparing similarities between region and text embeddings encoded by Contrastive Language-Image Pre-Training (CLIP)[14].

Grounding DINO[1] is an open-set object detector built on the DINO[15] architecture. It consists of an image encoder and a text encoder, followed by a transformer-based feature enhancer and cross-modality decoder to effectively align visual and textual modalities. Grounding DINO is pre-trained on large-scale detection and grounding datasets, including Objects365[16], Graphical Question Answering (GQA)[17], etc. It achieves a zero-shot performance of over 53.0 mAP on the COCO benchmark. Building upon Grounding DINO, Grounding DINO 1.5[2] further advances the model and becomes the state-of-the-art model by leveraging larger image backbones and over 20 million pre-training data. However, since Grounding DINO 1.5 is a closed-source model, we use Grounding DINO as an alternative in our experiments.

2.2 Few-Shot Object Detection

Few-shot object detection methods can generally be categorized into meta-learning-based approaches and transfer-learning-based approaches.

Meta-learning-based approaches learn class prototype representations for each base category and infer novel categories by aligning regions of interest (RoI) with prototype representations. For example, Meta Region-Based Convolutional Neural Networks (R-CNN)[18] conducted meta-learning over RoI features, using a support branch to create a category attention vector, which is then fused with RoI features for object detection. Ref. [19] leveraged LLMs to perform few-shot adaptation, relying on prototypes generated from DINO v2[20] and region proposals extracted using Deformable Detection Transformer (DETR)[21]. The two-stage fine-tuning approach (TFA)[22] proposed a two-phase fine-tuning method based on Faster R-CNN[23], surpassing prior meta-learning-based approaches by freezing the trained base class parameters and fine-tuning only the detection heads.

On the other hand, transfer-learning-based approaches focus on fine-tuning pre-trained object detectors using few-shot data. For example, Ref. [24] fine-tuned Grounding DINO to the agricultural domain in a few-shot setting by replacing the text encoder with randomly initialized trainable text embeddings.

CD-FSOD was initially proposed by Refs. [4] and [5]. They highlighted the importance of evaluating and enhancing the efficacy of pre-trained object detectors across varied domains that are rarely encountered in conventional internet-scale object detection pre-training datasets. Ref. [4] additionally outlined a couple of enhancements for adapting open-set object detectors (such as Grounding DINO and Detic[25]) for CD-FSOD, including prompt engineering, federated fine-tuning, and multi-modal prompting. Concurrently, Ref. [6] focused on optimizing data augmentation techniques when fine-tuning Grounding DINO for CD-FSOD.

Despite these efforts, to the best of our knowledge, no prior research has focused on developing a comprehensive end-to-end framework tailored for fine-tuning Grounding DINO specifically within CD-FSOD scenarios.

2.3 Mixture-of-Experts

The core concept of the MoE architecture is to enable different components (experts) of a model to specialize in different aspects of the data[10]. In recent years, MoE has gained significant popularity in LLMs, including Mistral[8] and DeepSeek-V2[9]. A common application of MoE in such models is to replace the traditional feed-forward network (FFN) with an MoE-based variant, as seen in Switch Transformer[26] and Open-MoE[27]. Each MoE layer consists of multiple ex-

perts, with only a subset activated for a given input. This selection is governed by a gating function (router), which dynamically routes different inputs to different experts. While MoE has been extensively explored in LLMs, to the best of our knowledge, we are the first to incorporate it into open-set object detection.

3 Method

3.1 End-to-End Framework for Cross-Domain Few-Shot Object Detection

In this work, we present a novel end-to-end framework that leverages the pre-trained Grounding DINO to address the challenge of data scarcity specifically in CD-FSOD. To the best of our knowledge, this is the first end-to-end framework for CD-FSOD. Our framework consists of three key stages, as conceptually illustrated in Fig. 1.

Firstly, an LLM-based text prompt optimization module is used. Given the raw category descriptions from the data, this module utilizes an LLM to generate more representative and effective text prompts, which are crucial for guiding Grounding DINO to better recognize novel concepts.

The second stage aims to adapt Grounding DINO to the target domain. In the CD-FSOD setting, since few bounding box annotations are available per category, not all target objects within an image are labeled. Therefore, we use a pseudo-labeling strategy to maximize the data utilization. This process generates additional training signals to enhance the fine-tuning performance of Grounding DINO.

In the last stage, we fine-tune our proposed MoE-Grounding DINO. This novel architecture is initialized from the Grounding DINO adapted in the second stage, and is further trained to substantially boost the overall model performance in the challenging CD-FSOD setting.

The detailed methodologies for each stage of this framework will be elaborated in subsequent sections.

3.2 LLM-Based Text Prompt Optimization

Grounding DINO leverages and aligns text and image modalities for object detection tasks. Its robust performance is attributed to pre-training on extensive internet-scale datasets, leading to a well-established alignment between text and image modalities. Therefore, when fine-tuning it on a downstream dataset, the quality of the text prompts (i.e., category descriptions here) becomes crucial. This is especially critical in CD-FSOD scenarios, where data scarcity limits the model's adaptation. In addition, an optimal alignment between the text prompts and Grounding DINO's pre-training data is also essential for maximizing the fine-tuning performance.

However, the quality of text prompts in downstream datasets is often a significant hurdle. For instance, in the actions-zzid2-zb1hq-fsod-amih dataset from the Roboflow20-VL[11] benchmark, the class "Attack" is described as "Players hit the ball over the net", and the class "Set" is described as "Players push the ball upwards with their fingertips". Such vague or overly simplistic prompts can make it difficult for Grounding DINO to learn robust representations, potentially leading to slow convergence or suboptimal performance during fine-tuning. Consequently, optimizing these text prompts before they are fed into Grounding DINO for fine-tuning is essential for achieving desirable results.

To address this, as illustrated in Fig. 2, we propose a three-step methodology that leverages an LLM, specifically Qwen2.5-VL[28], to optimize category descriptions. In the first step, Qwen2.5-VL is prompted with images from the dataset to generate concise scenario descriptions that capture the essence of the dataset effectively. In the second step, Qwen2.5-VL is tasked with generating detailed category descriptions for each class. In detail, given an image with a target object enclosed within a red bounding box, Qwen2.5-VL is prompted to produce multiple informative, representative and distinctive descriptions tailored to the target object based on the original category description and the dataset description generated in Step 1. In Step 3, we craft the ideal text prompt for each class. For each category, we fine-tune Grounding DINO with randomly combined textual descriptions generated in Stage 2, aiming to identify the most effective combination that serves as the optimal text prompt for that particular class.

3.3 MoE-Grounding DINO

In few-shot object detection, maximizing the object detector's ability to extract supervision from scarce data is crucial.

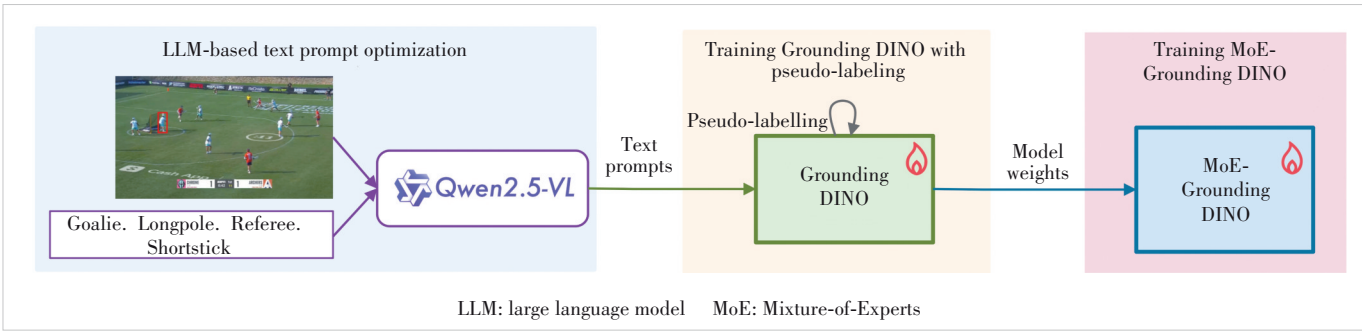

Figure 1. An illustration of the proposed end-to-end framework for cross-domain few-shot object detection (CD-FSOD)

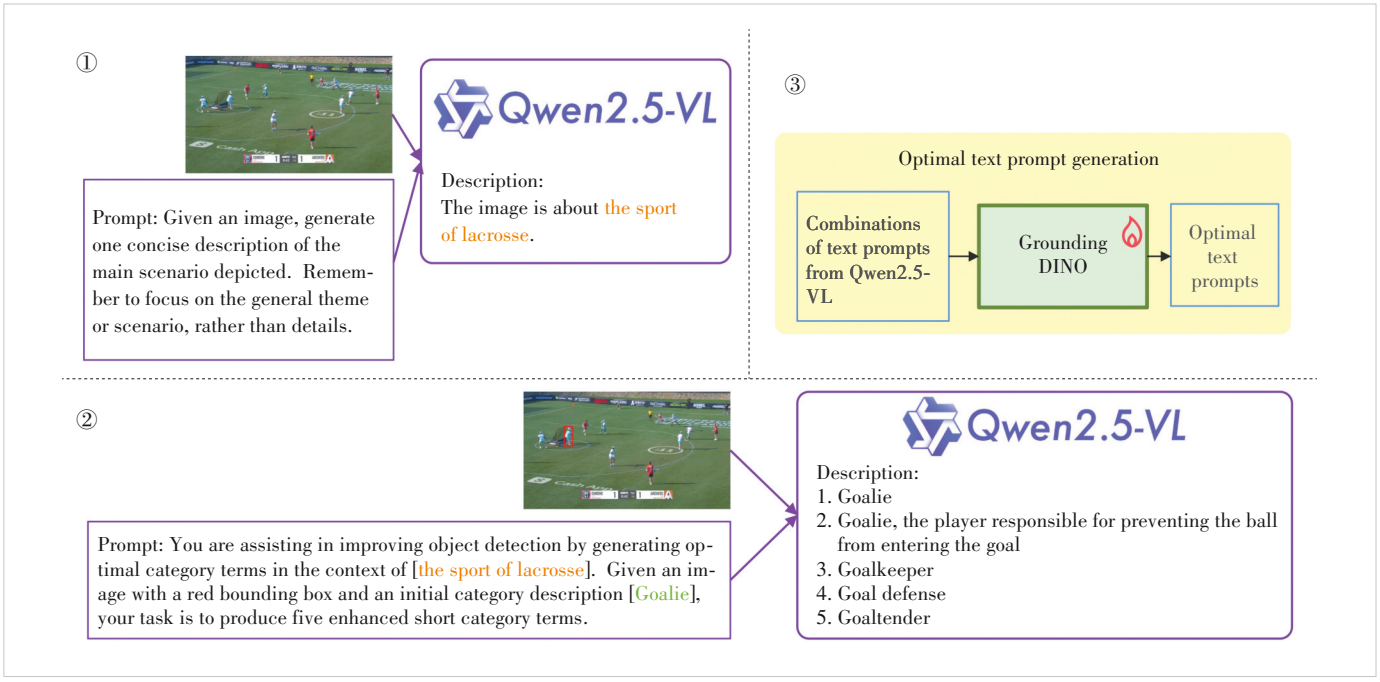

Figure 2. An example of the LLM-based text prompt optimization stage

In CD-FSOD, beyond the few-shot constraint, the cross-domain setting further increases the difficulty, as detectors usually require additional supervision to effectively adapt to a new domain. Therefore, success in CD-FSOD hinges on maximizing supervision and enabling the model to learn as much as possible from the limited training data.

The MoE architecture has demonstrated significant efficacy across LLMs. MoE functions by replicating individual experts multiple times and employs a dynamic routing mechanism. This mechanism enables different experts to specialize in distinct aspects of the input data, thereby empowering the model to capture more diverse and informative representations. It intuitively provides models with a broader parameter search space for learning, making it particularly well-suited for the CD-FSOD scenario.

We introduce MoE-Grounding DINO, a novel architecture built upon Grounding DINO. As depicted in Fig. 3, it incorporates the MoE module into Grounding DINO by substituting the FFN layers of the cross-modality decoder with MoE-FFN layers. The MoE-FFN layer, as detailed in Fig. 4, is designed to consist of a shared FFN and N routed FFNs (i.e., experts). For a given input, the router dynamically determines K routed FFNs to activate. The outputs of the shared FFN and K activated routed FFNs are then aggregated to form the output of the MoE-FFN layer. This process can be expressed as:

$$\begin{aligned}&\mathrm{MoE\_FFN}(x;\theta,\{\mathrm{FFN\_}r\}_{i=1}^{N},\mathrm{FFN\_}s)=\\&\mathrm{FFN\_}s(x)+\sum_{j=1}^{K}\vartheta(x;\theta)_j f_j(x;\mathrm{FFN\_}r_j)\end{aligned}\quad(1),$$

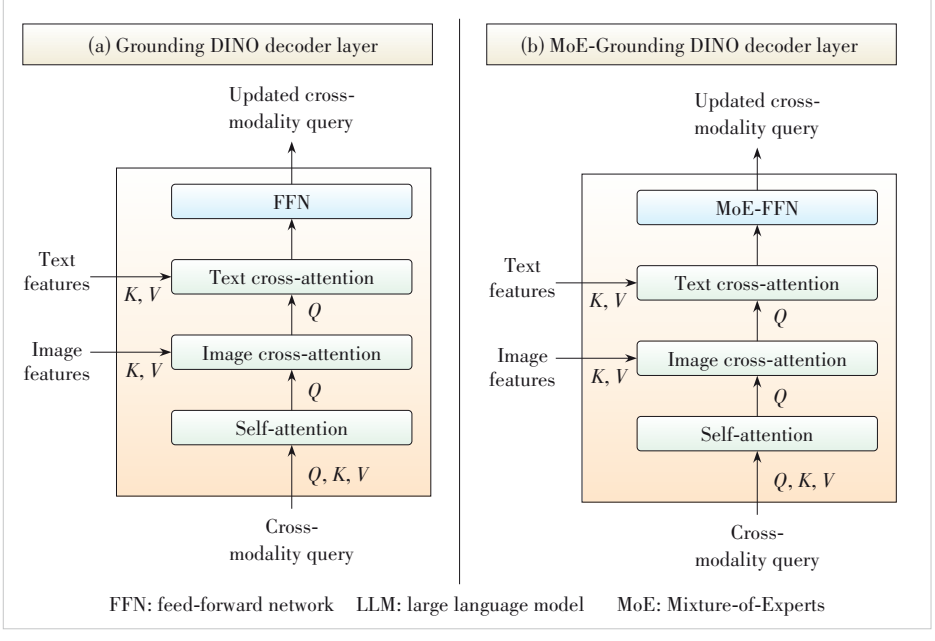

Figure 3. An illustration of the MoE-Grounding DINO decoder layer

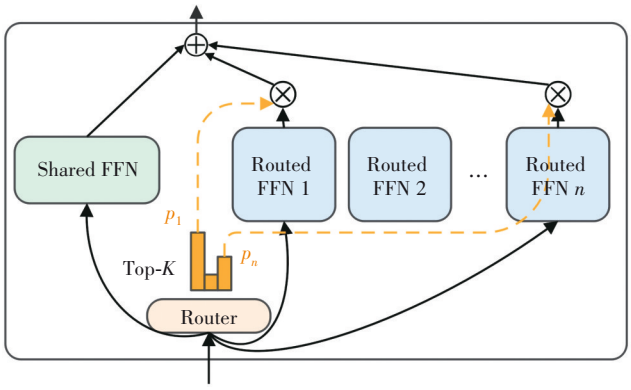

Figure 4. An illustration of the MoE-FFN layer

$$\vartheta(x;\theta)_j = \mathrm{softmax}(\mathrm{Top}K(g(x;\theta) + R_{\mathrm{noise}}, K))_j \quad (2),$$

$$f_j(x;\mathrm{FFN}\_r_j) = \mathrm{FFN}\_r_j(\mathrm{Dispatcher}(x)_j) \quad (3),$$

where FFN_r denotes individual routed FFNs, FFN_s refers to the shared FFN, $g(x;\theta)$ represents the routing function, R_{noise} is the noise term, and Dispatcher refers to the operation that dispatches the input to each expert.

Following Ref. [10], an additional auxiliary load balancing loss is applied during training. This loss is crucial for encouraging a more balanced utilization of the experts, preventing certain experts from dominating the model's learning.

The MoE-FFN module is seamlessly integrated into Grounding DINO without the need for retraining, as it can be initialized from the weights of Grounding DINO. This initialization process will be detailed later.

3.4 Training Strategy

Our proposed end-to-end framework (Fig. 3), is structured into three distinct stages: LLM-based text prompt optimization, Grounding DINO fine-tuning with pseudo-labeling, and MoE-Grounding DINO fine-tuning. This section elaborates on the methodologies and procedures employed for model training in Stages 2 and 3.

3.4.1 Grounding DINO Fine-Tuning with Pseudo-Labeling

The second stage is dedicated to fine-tuning the pre-trained Grounding DINO, augmented by a pseudo-labeling strategy. In CD-FSOD, only a limited number of bounding box annotations are typically provided per category. This often means that not all target objects within an image are initially labeled. To maximize the utilization of available data and compensate for annotation sparsity, we apply a simple pseudo-labeling approach to enhance Grounding DINO's performance.

As illustrated in Fig. 5, we first fine-tune Grounding DINO on the provided limited labeled training data. Subsequently, the fine-tuned model is employed to infer predictions on training instances. We only retain high-confidence predictions to ensure the quality of these inferred labels. This is followed by rigorous post-processing steps, such as Non-Maximum Suppression (NMS), to refine the generated pseudo-labels. The original ground-truth annotations are then combined with these high-confidence pseudo-labels to construct a refined training dataset. Finally, the fine-tuned Grounding DINO undergoes further training on this newly constructed refined dataset, leveraging the increased data volume to enhance its robustness and generalization.

3.4.2 MoE-Grounding DINO Fine-Tuning

The third stage involves fine-tuning our proposed MoE-Grounding DINO. For initialization, all components of MoE-Grounding DINO inherit the parameters from the Grounding

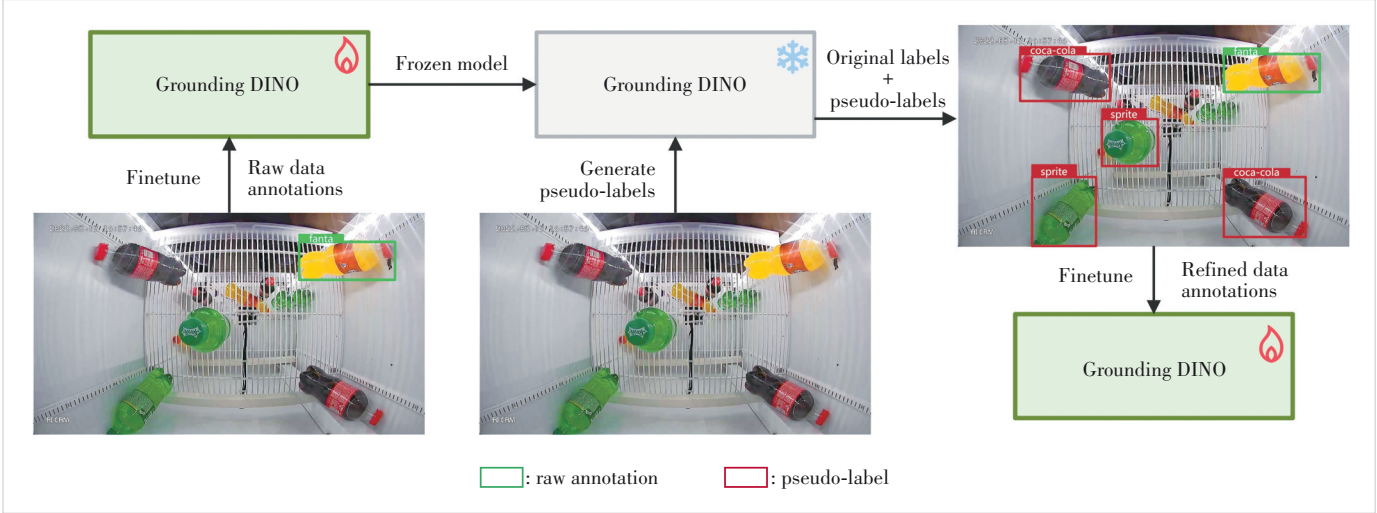

Figure 5. Illustration of the stage of fine-tuning Grounding DINO with pseudo-labeling refinement

DINO model fine-tuned in Stage 2, except for the newly introduced MoE-FFN layers. For these MoE-FFN layers, the shared FFN is initialized with the weights of the fine-tuned Grounding DINO to provide a solid foundation, while the routed FFNs are initialized with the weights from the pre-trained Grounding DINO model to encourage diversity among experts and facilitate robust training. The router, implemented as a simple FFN layer, is initialized randomly.

When fine-tuning MoE-Grounding DINO, we only allow the MoE-FFN layers, the bounding box regression head, and the classification head to be trainable to ensure stable training. As before, MoE-Grounding DINO is trained on the refined dataset constructed in Stage 2.

3.5 Data Augmentation

The success of object detectors in CD-FSOD hinges on learning robust and domain-invariant features from limited data. Therefore, we apply strong data augmentation techniques during fine-tuning to maximize the information extracted from constrained training data and to enhance feature diversity. Specifically, we combine RandomFlip, RandomResize, RandomCrop, YOLOXHSVRandomAug, and Copy-Paste to mitigate overfitting and enhance the variability of the training data.

4 Experiments

4.1 Datasets

We evaluate our approach on the Roboflow20-VL dataset[11], a widely adopted benchmark for CD-FSOD. This dataset is specifically designed to assess the generalization ability of object detectors on out-of-domain data, presenting a significant challenge as it contains concepts rarely encountered in internet-scale pre-training datasets. Roboflow20-VL comprises 20 datasets spanning seven diverse domains: Aerial, Document, Flora & Fauna, Industrial, Medical, Sports, and Others. To rigorously evaluate cross-domain adaptability, we adopt the provided 10-shot learning setup (i.e., only 10 bounding box annotations per class). Models are evaluated using the mAP metric independently for each class. We follow Ref. [29] for efficient dataset management.

4.2 Implementation Details

All experiments were conducted within the PyTorch framework on eight NVIDIA V100 GPUs. We used the Swin-L version of Grounding DINO with pre-trained weights from the mmdetection implementation (MM-GDINO-L*)[30], which represents the best-performing open-source model available. This model was pre-trained on large-scale datasets, including Objects365-V2[16], OpenImageV6[31], GoldG[32], COCO, and RefCOCO[33]. For LLM-based text prompt optimization, we used Qwen2.5-VL 7B. Our MoE-Grounding DINO was configured with N=3 experts and a top-K=2 routing mechanism.

During fine-tuning, we set the learning rate to 2^{-5} for Grounding DINO and 1^{-5} for MoE-Grounding DINO. Both models were fine-tuned for 24 epochs. All other training parameters (e.g., loss weights, number of queries) followed the default settings of Grounding DINO. For pseudo-label generation, we applied a confidence threshold of 0.5 and a Non-Maximum Suppression (NMS) Intersection over Union (IoU) threshold of 0.5. The relatively high confidence threshold was chosen empirically to prioritize label quality, as missing annotations only reduce data utilization, whereas incorrect annotations can significantly degrade model performance.

4.3 Main Results on Roboflow20-VL

Table 1 summarizes the CD-FSOD performance on the Roboflow20-VL benchmark, comparing our framework with TFA[22], Federated Detic[4], and Grounding DINO[1]. While the representative baseline TFA achieves competitive performance on in-domain few-shot benchmarks such as COCO, it struggles to generalize under the CD-FSOD setting (yielding only 9.8 mAP overall). Federated Detic, which employs federated fine-tuning to mitigate the impact of negative categories, performs better than TFA (20.3 mAP overall) but is still outperformed by directly fine-tuned Grounding DINO. Notably, directly fine-tuning Grounding DINO delivers a substantial improvement of 16.2 mAP over its zero-shot baseline (33.3 mAP vs. 17.1 mAP). However, this result remains far below its fine-tuning performance on the in-domain COCO dataset, which reaches 60.3 mAP as reported in Ref. [26]. This contrast underscores a crucial point: while fine-tuning Grounding DINO exhibits high performance on conventional object detection da-

Table 1. Comparison of performance of zero-shot, fine-tuned Grounding DINO with our proposed framework on Roboflow20-VL

Framework	Aerial	Document	Flora & Fauna	Industrial	Medical	Sports	Other	All
TFA[22]	9.4	3.8	16.8	14.4	2.7	1.3	10.2	9.8
Federated Detic[4]	11.6	14.3	30.8	24.7	8.9	17.4	21.0	20.3
Grounding DINO (zero-shot)	30.6	5.0	33.9	13.0	0.4	5.5	16.8	17.1
Grounding DINO (sft)	39.8	34.5	45.6	37.8	23.3	26.3	24.7	33.3
ETS[6]	41.6	27.4	48.1	49.2	27.4	30.9	33.7	36.9
Our framework	**49.6**	**46.3**	**55.0**	**61.3**	**42.5**	**41.2**	**45.1**	**48.7**

ETS: enhance then search TFA: two-stage fine-tuning approach

tasets like COCO, its performance under the CD-FSOD setting lags significantly behind expectations.

Table 1 clearly demonstrates that our proposed CD-FSOD framework notably outperforms directly fine-tuning Grounding DINO across all domains in the Roboflow20-VL benchmark. Specifically, our framework achieves an impressive overall enhancement of 15.4 mAP. In detail, our framework surpasses the baseline by 9.8 mAP, 11.8 mAP, 9.4 mAP, 23.5 mAP, 19.2 mAP, 14.9 mAP and 20.4 mAP in the Aerial, Document, Flora & Fauna, Industrial, Medical, Sports, and Others domains, respectively.

Furthermore, our work achieved first place in the Roboflow-20VL Few-Shot Object Detection Challenge[11] at the Workshop on Visual Perception and Learning in an Open World at CVPR 2025, where it was also prominently featured.

These results not only underscore the efficacy of our proposed framework, but also highlight its versatility and robustness across diverse domains under the CD-FSOD setting. The significant performance gains achieved through our methodology demonstrate its potential in enhancing object detection tasks in varied real-world scenarios.

4.4 Ablation Experiments

We conducted a series of ablation experiments to systematically evaluate the effectiveness of each component of our framework in the overall performance. Additionally, we explored the impact of the number of routed experts N in our proposed MoE-Grounding DINO architecture.

1) LLM-based text prompt optimization. As shown in Table 2, fine-tuning Grounding DINO with our optimized text prompts yields a significant improvement of 7.0 mAP compared to fine-tuning with the original dataset. This substantial gain underscores the critical role of high-quality text prompts in the CD-FSOD setting and validates the effectiveness of our proposed LLM-based text prompt optimization module in generating more descriptive and discriminative prompts.

2) Data augmentation strategies. Employing the data augmentation techniques we detailed before provides an additional benefit of 4.3 mAP (Table 2). This improvement highlights the importance of enhancing the diversity and variability of the training data, particularly in few-shot scenarios where data scarcity is a major challenge. By increasing data diversity, our augmentation strategies enable the model to generalize better to unseen instances.

3) Pseudo-labeling. After continuing to train Grounding DINO on a dataset augmented with pseudo-labels generated from an initial fine-tuned model, we observe a further performance improvement of 1.9 mAP (Table 2). This demonstrates the effectiveness of our pseudo-labeling approach in maximizing data utilization under the CD-FSOD setting, allowing the model to learn from a larger and more comprehensive set of examples, thereby mitigating the limitations of limited labeled data.

4) Fine-tuning MoE-Grounding-DINO. We evaluate the impact of fine-tuning our proposed MoE-Grounding DINO. Building upon the previously fine-tuned Grounding DINO, incorporating the MoE modules further boosts the overall detection performance by 2.2 mAP (Table 2). This significant improvement highlights the effectiveness of employing the MoE module to seamlessly expand the parameter space of Grounding DINO, thereby handling the data scarcity challenge present in the CD-FSOD context.

5) Effect of the number of routed experts. Using three routed experts yields the optimal performance, as shown in Fig. 6. Using a larger number of experts provides no addi-

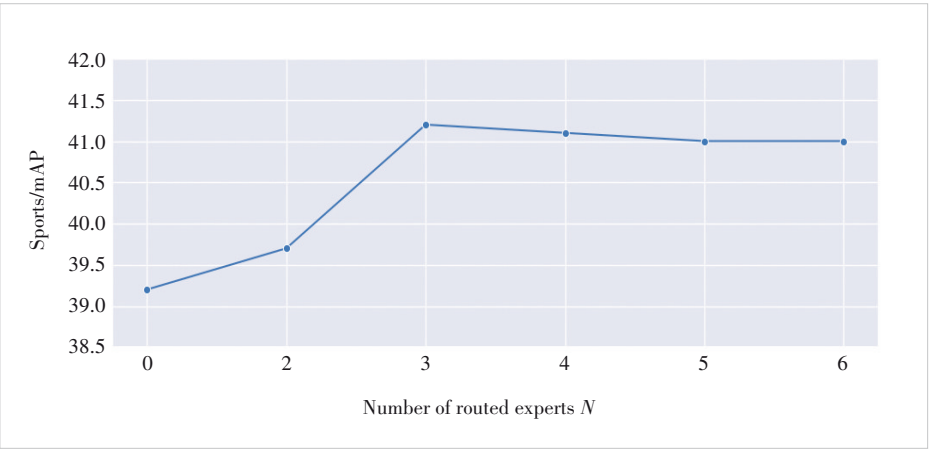

Figure 6. Effect of the total number of routed experts (N) in MoE-Grounding DINO, evaluated on the Sports domain, with K fixed at 2. An N value of 0 indicates the baseline model without the MoE module

Table 2. Effectiveness of different components in our end-to-end framework on Roboflow20-VL

Component	Aerial	Document	Flora & Fauna	Industrial	Medical	Sports	Other	All
Grounding DINO (sft)	39.8	34.5	45.6	37.8	23.3	26.3	24.7	33.3
+ text prompt optimization	39.9	40.2	49.7	47.4	36.2	33.8	34.8	40.3 (**+7.0**)
+ data augmentation	48.2	42.5	53.6	52.0	37.8	38.1	40.2	44.6 (**+4.3**)
+ pseudo-labeling	**49.6**	44.0	54.6	56.4	40.2	39.2	41.6	46.5 (**+1.9**)
+ MoE-Grounding DINO	**49.6**	**46.3**	**55.0**	**61.3**	**42.5**	**41.2**	**45.1**	**48.7** (+2.2)

tional benefit. We attribute this to the few-shot learning scenario, where the limited number of training samples may be insufficient to effectively fine-tune a larger number of experts.

5 Conclusions

In this paper, we introduce the first end-to-end framework for cross-domain few-shot object detection, which significantly advances the state of the art on the Roboflow20-VL benchmark. Our framework uniquely integrates LLM-based text prompt optimization and a multi-stage training pipeline with pseudo-labeling. Another core contribution of our work is the novel MoE-Grounding DINO, which marks the pioneering application of the mixture-of-experts in open-set object detection.

Although representing a significant advancement, our framework still faces certain limitations. Firstly, the pseudo-labeling effectiveness can be sensitive to the initial model's performance and the severity of the domain shift. Noise in pseudo-labels can potentially propagate errors. Secondly, the substitution of MoE modules introduces computational and memory overhead, challenging real-time deployment and resource-constrained devices.

These limitations require further attention and improvement in future research to enhance the robustness and applicability of our framework. In addition, we will integrate the MoE architecture into other open-set object detectors to validate its efficacy and generalizability beyond Grounding DINO.

References

[1] LIU S L, ZENG Z Y, REN T H, et al. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection [C]//European Conference on Computer Vision. ECCV, 2024: 38 – 55. DOI: 10.1007/978-3-031-72970-6_3
[2] REN T H, JIANG Q, LIU S L, et al. Grounding DINO 1.5: advance the "edge" of open-set object detection [EB/OL]. [2024-06-01]. https://arxiv.org/abs/2405.10300
[3] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]//European Conference on Computer Vision. ECCV, 2014: 740 – 755. DOI: 10.1007/978-3-319-10602-1_48
[4] MADAN A, PERI N, KONG S, et al. Revisiting few-shot object detection with vision-language models [C]//Proc. Advances in Neural Information Processing Systems 37. NeurIPS, 2024: 19547 – 19560. DOI: 10.52202/079017-0617
[5] FU Y Q, WANG Y, PAN Y X, et al. Cross-domain few-shot object detection via enhanced open-set object detector [C]//European Conference on Computer Vision. ECCV, 2024: 247 – 264. DOI: 10.1007/978-3-031-73636-0_15
[6] PAN J C, LIU Y X, HE X, et al. Enhance then search: an augmentation-search strategy with foundation models for cross-domain few-shot object detection [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2025: 1539 – 1547. DOI: 10.1109/CVPRW67362.2025.00143
[7] JACOBS R A, JORDAN M I, NOWLAN S J, et al. Adaptive mixtures of local experts [J]. Neural computation, 1991, 3(1): 79 – 87
[8] JIANG A Q, SABLAYROLLES A, ROUX A, et al. Mixtral of experts [R]. 2024
[9] LIU A, FENG B, WANG B, et al. Deepseek-v2: a strong, economical, and efficient mixture-of-experts language model [R]. 2024
[10] CAI W, JIANG J, WANG F, et al. A survey on mixture of experts [R]. 2024
[11] ROBICHEAUX P, POPOV M, MADAN A, et al. Roboflow100-vl: a multi-domain object detection benchmark for vision-language models [R]. 2025
[12] LI L H, ZHANG P C, ZHANG H T, et al. Grounded language-image pre-training [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 10955 – 10965. DOI: 10.1109/CVPR52688.2022.01069
[13] ZHONG Y W, YANG J W, ZHANG P C, et al. RegionCLIP: region-based language-image pretraining [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 16772 – 16782. DOI: 10.1109/CVPR52688.2022.01629
[14] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning. PMLR, 2021: 8748 – 8763
[15] ZHANG H, LI F, LIU S, et al. Dino: DETR with improved denoising anchor boxes for end-to-end object detection [R]. 2022
[16] SHAO S, LI Z M, ZHANG T Y, et al. Objects365: a large-scale, high-quality dataset for object detection [C]//Proc. IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 8430 – 8439. DOI: 10.1109/iccv.2019.00852
[17] HUDSON D A, MANNING C D. GQA: a new dataset for real-world visual reasoning and compositional question answering [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 6693 – 6702. DOI: 10.1109/CVPR.2019.00686
[18] YAN X P, CHEN Z L, XU A N, et al. Meta R-CNN: towards general solver for instance-level low-shot learning [C]//Proc. IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 9577 – 9586. DOI: 10.1109/iccv.2019.00967
[19] HAN G X, LIM S N. Few-shot object detection with foundation models [C]//Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024: 28608 – 28618. DOI: 10.1109/CVPR52733.2024.02703
[20] OQUAB M, DARCET T, MOUTAKANNI T, et al. Dinov2: learning robust visual features without supervision [R]. 2023
[21] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection [R]. 2021
[22] WANG X, HUANG T, GONZALEZ J, et al. Frustratingly simple few-shot object detection [C]//International Conference on Machine Learning. PMLR, 2020: 9919 – 9928
[23] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]//IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137 – 1149. DOI: 10.1109/TPAMI.2016.2577031
[24] SINGH R, PUHL R B, DHAKAL K, et al. Few-shot adaptation of grounding DINO for agricultural domain [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2025: 5332 – 5342. DOI: 10.1109/CVPRW67362.2025.00530
[25] ZHOU X, GIRDHAR R, JOULIN A, et al. Detecting twenty-thousand classes using image-level supervision [C]//European Conference on Computer Vision. ECCV, 2022: 350 – 368. DOI: 10.1007/978-3-031-20077-9_21
[26] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity [J]. Journal of machine learning research, 2022, 23(120): 1 – 39
[27] XUE F Z, ZHENG Z A, FU Y, et al. OpenMoE: an early effort on open mixture-of-experts language models [R]. 2024
[28] BAI S, CHEN K Q, LIU X J, et al. Qwen2.5-VL technical report [R]. 2025
[29] HAN Y J, NIU J H, TU Y F. Development trends and challenges of data management systems [J]. ZTE technology journal, 2023, 27(2): 64 – 71. DOI: 10.12142/ZTETJ.202304012
[30] ZHAO X Y, CHEN Y C, XU S L, et al. An open and comprehensive pipeline for unified object grounding and detection [R]. 2024

[31] KUZNETSOVA A, ROM H, ALLDRIN N, et al. The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale [J]. International journal of computer vision, 2020, 128 (7): 1956 – 1981. DOI: 10.1007/s11263-020-01316-z
[32] JENKINS P, SACHDEVA R, KEBE G Y, et al. Presentation and analysis of a multimodal dataset for grounded language learning [R]. 2020
[33] KAZEMZADEH S, ORDONEZ V, MATTEN M, et al. ReferItGame: referring to objects in photographs of natural scenes [C]//Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). USAACL, 2014: 787 – 798. DOI: 10.3115/v1/d14-1086

Biographies

DONG Xiugang is an AI engineer at ZTE Corporation. His work primarily focuses on the research and development of large computer vision and video understanding models. His research interests span multiple areas, including open-set object detection, semantic segmentation, video spatio-temporal localization, and general video understanding.

ZHANG Kaijin is an AI engineer at ZTE Corporation, specializing in the research and development of large computer vision models. His research interests include open-set object detection, semantic segmentation, and keypoint detection.

NONG Qingpeng is an AI engineer at ZTE Corporation, specializing in the research and development of large computer vision models. His research interests include open-set object detection, semantic segmentation, and keypoint detection.

JU Minhan received his Bachelor's degree in data science and big data technology from Xi'an Jiaotong-Liverpool University, China. He is currently pursuing a master's degree in data science at the University of Sydney, Australia. His research interests include machine learning, data mining, and big data system modeling. He is now interning at ZTE Corporation.

TU Yaofeng (tu.yaofeng@zte.com.cn) is the Deputy Dean of the Central Research Institute of ZTE Corporation. As a PhD and senior researcher, he focuses his research on big data, databases, AI, large models, and cloud computing.