Simple Technical Report for the Foundational Few-Shot Object Detection Challenge v2 2025

Kaijin Zhang[1], Xuezhen Tu[2], Qingpeng Nong[1], Xiugang Dong[1], Xurui Gao[1], Xiangsheng Zhou[1]
[1]Central R & D Institute, ZTE
[2]Shanghai Jiao Tong University

Abstract

Open-set object detectors, such as Grounding DINO, have exhibited impressive zero-shot detection performance on common benchmarks like COCO and LVIS. However, they often encounter challenges when detecting objects in specialized domains which are not well-represented in their millions of training data, such as medical and aerial imagery.

The Foundational Few Shot Object Detection Challenge v2 focuses on the adaptation of pre-trained object detectors to specific target domains using only a small number of examples per class (specifically 10-shot in this context). This report provides a concise overview of our approach to addressing the aforementioned challenge, highlighting our methodology and strategies employed.

1. Foundation Model

1.1. Model Description

We use Nebula-CV as our foundational model. As Nebula-CV is an unpublished open-set object detection model, we provide only a succinct overview of its key attributes here. We present the overall model architecture of Nebula-CV in Figure 1.

Nebula-CV is built on the DINO [9] architecture. It achieves open-set object detection by fusing the text modality into the model. The fusion of both modalities is facilitated by carefully designed attention modules. Nebula-CV leverages Swin-B [4] as its visual backbone and BERT [1] as the text encoder.

1.2. Pre-training and Pre-training Data

Nebula-CV follows a two-stage pre-training paradigm. In the first stage, Nebula-CV undergoes pre-training on around 5M carefully curated and refined web-scale data. In the second stage, the model is further fine-tuned on around 1M high-quality grounding data distilled from Qwen2.5-VL [8].

2. Few-shot Domain Adaptation

For this challenge, our work on few-shot domain adaptation can be categorized into four main components: text prompt optimization, data augmentation optimization, pseudo-label optimization, and inference resolution optimization.

2.1. Text Prompt Optimization

A subset of Roboflow-VL [7] is used in this challenge. However, upon careful investigation, the provided category names are often vague, lack distinguishing features, and misalign with the pre-training data.

To address this issue, We leverage Qwen2.5-VL [8] for the generation of category descriptions for each dataset, employing a two-stage methodology:

stage 1: Given images from a dataset, Qwen2.5-VL is tasked with producing concise descriptions of the information present in the dataset images.

stage 2: For each dataset, Qwen2.5-VL is prompted to generate descriptive category names for each class. An example prompt could be:

"You are assisting in improving object detection by generating optimal category terms in the context of [descriptions generated in stage 1]. Given an image with a red bounding box and an initial category description [original class description], your task is to produce five enhanced short category terms."

Subsequently, we randomly combine the text prompts generated in Stage 2 to derive the optimal text descriptions for each class.

2.2. Data Augmentation Optimization

Our experiments highlight the pivotal role of data augmentation in few-shot object detection. These techniques not only enhance sample diversity but also simulate domain shifts effectively. Through our experimentation, we have found that combinations of RandomFlip, RandomResize, RandomCrop, YOLOXHSVRandomAug, and Copy-Paste consistently yield satisfactory results. While we have also

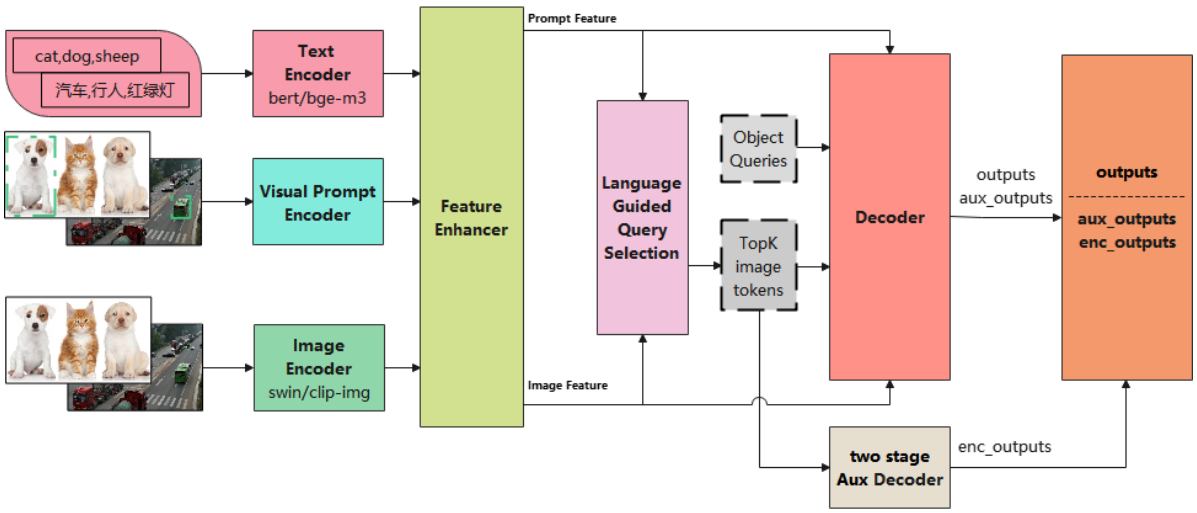

Figure 1. The overall model architecture of Nebula-CV.

explored other methods such as CachedMixUp, they did not yield significant benefits in the few-shot setting.

2.3. Pseudo-label Optimization

The Foundational Few-Shot Object Detection Challenge v2 provides only 10-shot examples per class as raw training data, resulting in sparse annotations. To mitigate this, we adopt a pseudo-labeling pipeline: (1) train initial models on the raw data, (2) infer labels for training instances, (3) apply rigorous post-processing steps to refine pseudo-labels, and (4) retrain models on the augmented dataset to further boost performance.

2.4. Inference Resolution Optimization

While models like Grounding DINO [3] and Grounding DINO 1.5 [6] typically utilize an inference resolution of 800*1333, a fixed resolution may not be optimal for all downstream datasets. To tackle this challenge, we determine the most suitable inference resolution based on the training data resolutions for each dataset. Our research indicates that this optimization technique enhances model performance.

2.5. Other Details

We also experimented with additional techniques proven effective in previous work, including federated fine-tuning [5] and LLM-based post-processing [2]. However, our results indicate that these methods did not produce measurable improvements for our model.

3. Conclusion

This report succinctly outlines our approach to the Foundational Few-Shot Object Detection Challenge v2. We initially introduce the foundational model alongside its pre-training strategy. Subsequently, we detail our methodologies tailored to this challenge, focusing on few-shot domain adaptation.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 1

[2] Yuqian Fu, Xingyu Qiu, Bin Ren, Yanwei Fu, Radu Timofte, Nicu Sebe, Ming-Hsuan Yang, Luc Van Gool, Kaijin Zhang, Qingpeng Nong, et al. Ntire 2025 challenge on cross-domain few-shot object detection: Methods and results. *arXiv preprint arXiv:2504.10685*, 2025. 2

[3] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[5] Anish Madan, Neehar Peri, Shu Kong, and Deva Ramanan. Revisiting few-shot object detection with vision-language models. *Advances in Neural Information Processing Systems*, 37:19547–19560, 2024. 2

[6] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 2

[7] Peter Robicheaux, Matvei Popov, Anish Madan, Isaac Robinson, Joseph Nelson, Deva Ramanan, and Neehar Peri. Roboflow100-vl: A multi-domain object detection benchmark for vision-language models. *arXiv preprint arXiv:2505.20612*, 2025. 1

[8] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[9] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1