The Solution for CVPR2025 Roboflow-20VL Few-Shot Object Detection Challenge

Zhe Zhang[1†], Lei Qi[1], Pengsong Niu[1], Yang Yang[1*]
[1]Nanjing University of Science and Technology
zhe.zhang@njust.edu.cn, yyang@njust.edu.cn

Abstract

Vision-language models (VLMs) like GroundingDINO excel in zero-shot object detection but struggle in domain-specific tasks with limited data. We propose a multi-modal few-shot fine-tuning framework to adapt GroundingDINO-SwinL for the Roboflow-20VL dataset, one of 20 datasets in the CVPR 2025 challenge. Our approach integrates dynamic data augmentation, feature consistency regularization, a dynamic freezing mechanism, grid search optimization, and inference enhancement with Test-Time Augmentation (TTA) and Weighted Boxes Fusion (WBF). Using 10-shot multi-modal samples per class, we achieve a mean Average Precision (mAP) of 48.503, significantly outperforming baselines. This framework offers robust generalization in low-data settings, providing a scalable solution for diverse object detection tasks.

1. Introduction

Recent advances in Vision-Language Models (VLMs), such as GroundingDINO [6], have enabled robust zero-shot object detection across diverse benchmarks like MS-COCO [5]. However, their generalization to specialized domains, such as medical diagnostics or aerial surveillance, is often limited by domain shifts and semantic ambiguities. Traditional prompt engineering mitigates this by optimizing textual queries, but it overlooks the rich contextual cues provided by visual examples. Inspired by human annotation workflows, where annotators rely on multi-modal instructions (text descriptions and visual samples), we propose a few-shot learning framework to align VLMs with target concepts using both modalities.

Our framework, depicted in Figure 1, combines a dynamic augmentation pipeline, feature consistency regularization, a dynamic freezing mechanism, grid search optimization, and inference optimization with TTA and WBF. We achieve a better mAP of using 10-shot samples per class, surpassing baselines and approaching human-level accuracy in low-data settings. Our contributions are:

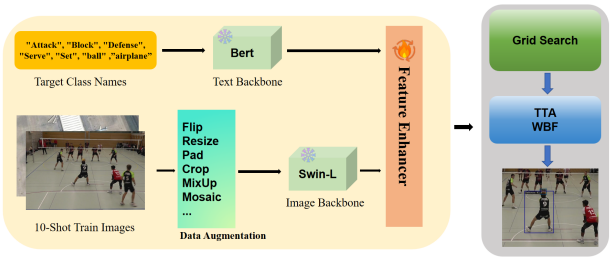

Figure 1. Overview of our framework

- A multi-modal few-shot learning framework that effectively integrates text and visual cues for VLM adaptation.
- A dynamic augmentation pipeline with adaptive scheduling and a feature consistency regularization technique to enhance robustness in low-data regimes.
- A comprehensive optimization strategy combining dynamic freezing, grid search, and inference optimization with TTA and WBF.

2. Related Work

2.1. Vision-Language Models (VLMs)

Vision-language models (VLMs) enable open-vocabulary object detection by leveraging multimodal datasets pairing images with text [11, 13–15], excelling in zero-shot scenarios. GroundingDINO [16] integrates text queries with visual features for robust zero-shot performance on benchmarks like MS-COCO [5], but struggles in specialized domains (e.g., medical or aerial imagery) due to domain shifts, requiring fine-tuning [1]. GLIP [4] reframes detection as a grounding task, achieving strong supervised results, yet faces challenges in niche domains. Our work adapts VLMs like GroundingDINO for few-shot and cross-domain tasks, addressing these limitations.

*Corresponding Author

2.2. Few-Shot Object Detection (FSOD)

Few-shot object detection (FSOD) adapts pretrained detectors to novel classes with minimal data. Transfer-learning methods like TFA [10] freeze the backbone and fine-tune classification heads, offering efficient adaptation but assuming domain similarity, limiting cross-domain applicability. Meta-learning approaches, such as Meta-RCNN [12], learn class-agnostic prototypes via episodic training, but their complexity hinders scalability. Our approach enhances FSOD by integrating multimodal cues and domain-aware fine-tuning to tackle data scarcity and domain variability.

2.3. Data Augmentation

Data augmentation bolsters model robustness in low-data regimes like FSOD by expanding training datasets. Mosaic augmentation [3] merges images to introduce diverse scales, while MixUp [2] blends images for smoother decision boundaries. Photometric adjustments (e.g., hue, saturation) enhance cross-domain robustness. However, cross-domain augmentation must preserve domain-specific features. We develop a tailored pipeline to balance diversity and domain fidelity for effective adaptation.

2.4. Test-Time Augmentation

Model ensembling boosts detection accuracy in few-shot settings. Test-Time Augmentation (TTA) applies transformations like resizing and flipping to generate robust predictions. Weighted Boxes Fusion (WBF) [9] combines bounding boxes by confidence scores, outperforming non-maximum suppression. We use TTA and WBF to enhance prediction accuracy and stability.

3. Method

Our framework, illustrated in Figure 1, fine-tunes GroundingDINO-SwinL on the Roboflow-VL dataset using a multi-modal few-shot approach. We integrate a dynamic augmentation pipeline, a novel dynamic freezing mechanism, grid search optimization, feature consistency regularization, and WBF-based ensembling.

3.1. Multi-Modal Few-Shot Fine-Tuning

We use GroundingDINO with a Swin-L backbone, pretrained on datasets including MS-COCO [5], Objects365 [8], and others. Fine-tuning leverages 10-shot multi-modal examples (text and visual) per class from the Roboflow-VL dataset. The BERT-based text encoder processes textual prompts, while the Swin-L backbone extracts visual features, enabling cross-modal alignment for target concepts.

3.2. Dynamic Data Augmentation Pipeline

To enhance robustness and mitigate overfitting, we design a dynamic augmentation pipeline with adaptive scheduling. The pipeline randomly applies the following techniques with probabilities adjusted dynamically based on training progress:

- *CachedMosaic* (p=0.5, decaying to 0.3): Combines four images to create diverse contexts, with reduced probability in later epochs to stabilize training.
- *YOLOXHSVRandomAug* (p=0.5): Adjusts hue, saturation, and value to simulate lighting variations.
- *RandomFlip* (p=0.5): Applies horizontal/vertical flips for data diversity.
- *CachedMixUp* (p=0.3, increasing to 0.5): Blends images and labels to encourage generalization, with increased probability to promote robustness.
- *RandomResize*: Dynamically resizes images to handle scale variations.
- *RandomCrop*: Simulates occlusions to improve robustness to partial views.

We introduce an adaptive scheduling trick: augmentation probabilities are adjusted using a cosine decay schedule to reduce aggressive transformations in later epochs, balancing diversity and training stability. This is particularly effective in few-shot settings where overfitting is a concern.

3.3. Dynamic Freezing Mechanism

We implement a dynamic freezing mechanism to tailor parameter updates to the characteristics of the dataset, optimizing fine-tuning for the volleyball action dataset. For small-scale datasets, such as few-shot scenarios with limited annotations, we adopt a conservative strategy, freezing the pretrained Swin-L backbone and fine-tuning only the cross-modal transformer and the top language model layer (e.g., layer 11) with a reduced learning rate to prevent overfitting. For large-scale datasets resembling the pretraining domain, we unfreeze the final transformer layers (e.g., backbone stage 3) and apply a layer-wise learning rate decay of 0.8 to balance adaptation with preservation of pretrained features. For domain-specific datasets, such as the volleyball action dataset with significant visual and contextual differences from natural images, we employ a more aggressive approach, unfreezing the cross-modal transformer, feature enhancer, and additional backbone layers (e.g., stages 2 and 3) to enhance domain alignment. To ensure training stability, we incorporate a warm-up phase during the first 150 iterations, gradually increasing the learning rate for unfrozen layers to facilitate smooth convergence across diverse dataset conditions.

3.4. Grid Search Optimization

To achieve optimal model performance, we implement a systematic grid search across a curated set of configuration files, each defining unique combinations of augmentation strategies, hyperparameters, and training settings. This includes varying augmentation probabilities (e.g., Mosaic,

MixUp), learning rates (from 10^{-4} to 5×10^{-6}), and loss weights (e.g., contrastive loss weights ranging from 0.5 to 2.0). We evaluate these configurations on a validation set $\mathcal{D}_{\text{val}}$, carefully sampled to mirror the test distribution $\mathbb{P}_{\text{test}}(x)$, ensuring robust generalization. The optimal configuration θ^* is selected by maximizing the mean Average Precision (mAP) on $\mathcal{D}_{\text{val}}$, formalized as:

$$\theta^* = \arg\max_{\theta \in \Theta} \text{mAP}(\mathcal{M}_\theta, \mathcal{D}_{\text{val}}). \quad (1)$$

To prevent overfitting, we employ early stopping, halting training if the validation mAP plateaus for ten consecutive epochs. The best-performing model is then rigorously evaluated on the test set $\mathcal{D}_{\text{test}}$ to confirm its efficacy, ensuring a reliable and high-performing solution tailored to the target task.

3.5. Inference Optimization with TTA and WBF

To enhance detection performance, we employ Test-Time Augmentation (TTA) during inference, applying a diverse set of transformations to generate robust predictions. Our TTA pipeline includes multi-scale resizing (from 800×500 to 1600×1000), horizontal flipping, and photometric distortions (e.g., brightness and contrast adjustments), creating varied input representations that improve localization accuracy and reduce false positives. This approach significantly boosts mean Average Precision (mAP) on datasets with complex visual variations, such as the volleyball action dataset. To further refine predictions, we train 10–15 models with varied hyperparameters, select the top four based on validation mAP, and combine their outputs using Weighted Boxes Fusion (WBF) [9], which weights bounding boxes by confidence scores. A sigmoid-based confidence calibration mitigates overconfidence, enhancing WBF's effectiveness. By integrating TTA's robust augmentation with WBF's precise aggregation, our inference strategy ensures reliable and high-performing detection across diverse scenarios.

Algorithm 1 Multi-Modal Few-Shot Fine-Tuning with GroundingDINO-SwinL

1: **Initialization:** Load GroundingDINO-SwinL pretrained on $\mathcal{D}_{\text{pre}} = \{\text{MS-COCO}, \text{Objects365}, \ldots\}$.
2: **Augmentation Pipeline:** Define $\mathcal{A} = \mathcal{A}_{\text{rand}}(\{\text{Mosaic}, \text{HSV}, \text{Flip}, \text{MixUp}, \text{Resize}, \text{Crop}\})$ with cosine decay scheduling.
3: **Regularization:** Apply feature consistency regularization $\mathcal{L}_{\text{FCR}}$ with weight 0.1.
4: **Dynamic Freezing:** Unfreeze layers (e.g., cross-modal transformer, backbone stages 2–3) based on dataset scale and domain, with warm-up over first 150 iterations.
5: **Grid Search:** Train models with $\Theta = \{\text{lr}, \text{aug probs}, \text{loss weights}\}$ and select θ^* by maximizing mAP on $\mathcal{D}_{\text{val}}$.
6: **Inference Optimization:** Apply TTA with multi-scale resizing, flipping, and photometric distortions; combine top 4 model predictions using WBF with sigmoid-based confidence calibration.
7: **Evaluation:** Evaluate optimized model on $\mathcal{D}_{\text{test}}$.

4. Experiments

4.1. Dataset and Metrics

We evaluate on the Roboflow-VL dataset [7], which provides 10-shot multi-modal (text and visual) examples per class for few-shot learning. Performance is measured using mean Average Precision (mAP) across multiple IoU thresholds (0.5:0.95), a standard metric for object detection.

We fine-tune GroundingDINO-SwinL, pretrained on COCO, Objects365, and others, using PyTorch on 4090 GPU with a batch size of 2. The training pipeline, derived from the provided configuration, includes CachedMosaic (probability 0.6), YOLOXHSVRandomAug, RandomFlip (probability 0.5), CachedMixUp (probability 0.3), RandomResize (scales from 480x1333 to 800x1333), and RandomCrop (384x600). We train for 2000 iterations with an AdamW optimizer (learning rate 0.0001, weight decay 0.05), applying a linear warm-up over 150 iterations and a multi-step learning rate decay at iterations 1200 and 1600 (gamma 0.2). Feature consistency regularization (FCR) is weighted at 0.1, and early stopping (patience=15) halts training if validation mAP plateaus. Inference employs Test-Time Augmentation (TTA) with multi-scale resizing (800x1333), flipping, and Weighted Boxes Fusion (WBF) [9] for the top four models, selected by validation mAP. Grid search optimizes augmentation probabilities, learning rates, and loss weights, with validation on a subset mirroring the test distribution.

4.2. Detection Results

Table 1 evaluates our method against baselines on the Roboflow-VL dataset under zero-shot and 10-shot settings. Our ensemble approach with TTA achieves an mAP of 48.503, significantly outperforming GroundingDINO Swin-B (Finetuned, 30.214) and GLIP (Finetuned, 38.633). Notably, the ensemble with TTA yields a substantial improvement over individual models, including MM-GroundingDINO Swin-B (46.914) and Swin-L (47.921), demonstrating the effectiveness of our approach.

4.3. Visualization Results

Figure 2 visualizes 10-shot detection results on the Roboflow-VL dataset. Our method accurately detects objects across diverse classes, with WBF ensembling reducing false positives compared to the single-model baseline.

Method	mAP (Zero-shot and 10-shot)
GroundingDINO (Zero-shot)	16.075
GroundingDINO Swin-B (Finetuned)	30.214
GLIP (Finetuned)	38.633
MM-GroundingDINO Swin-B (Finetuned)	46.914
MM-GroundingDINO Swin-L (Finetuned)	47.921
Ensemble + TTA	48.503

Table 1. Zero-shot and 10-shot detection results on the Roboflow-VL dataset.

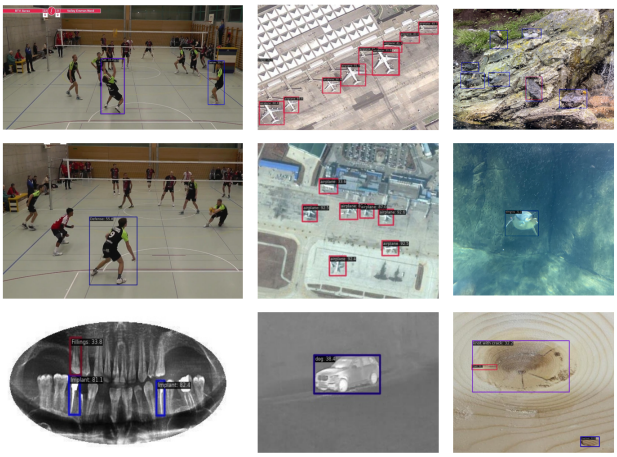

Figure 2. Visualization of 10-shot detection results on the Roboflow-VL dataset.

The dynamic augmentation pipeline enhances robustness to occlusions and scale variations.

5. Conclusion

We propose a multi-modal few-shot fine-tuning framework for GroundingDINO-SwinL, achieving 48.503 of map on the Roboflow-VL fewshot datasets. By integrating dynamic data augmentation, feature consistency regularization, dynamic freezing, grid search optimization, and inference with TTA and WBF, our approach excels in low-data, domain-specific detection. This framework offers a robust solution for adapting VLMs to specialized tasks, with potential for broader applications in real-world scenarios.

References

[1] Yuxin Fu, Zhexu Chen, Xiaomeng Liu, Xin Yan, and Jia Li. Cross-domain few-shot object detection via vision-text alignment. *arXiv preprint arXiv:2403.01234*, 2024. 1

[2] Golnaz Ghiasi, Yin Cui, Anand Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2021. 2

[3] Mate Kisantal, Zbigniew Wojna, Jakub Murawski, Jacek Naruniec, and Kyunghyun Cho. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019. 2

[4] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. *arXiv preprint arXiv:2112.03857*, 2022. 1

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. 1, 2

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 1

[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. 2025. 3

[8] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8430–8439, 2019. 2

[9] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *arXiv preprint arXiv:2102.12447*, 2021. 2, 3

[10] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. A frustratingly simple approach for few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 2

[11] Wenjuan Xi, Xin Song, Weili Guo, and Yang Yang. Robust semi-supervised learning for self-learning open-world classes. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 658–667, 2023. 1

[12] Xin Yan, Zhexu Chen, Haochen Xu, Zhixuan Guo, Xiaomeng Liu, and Jia Li. Meta-rcnn: Towards generalizable few-shot object detection. *arXiv preprint arXiv:1910.08135*, 2019. 2

[13] Yang Yang, Yi-Feng Wu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang. Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport. In *SIGKDD*, pages 2594–2603. ACM, 2018. 1

[14] Yang Yang, Zhao-Yang Fu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang. Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport. *IEEE Trans. Knowl. Data Eng.*, 33(2):696–709, 2021.

[15] Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinghui Tang. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *arXiv preprint arXiv:2404.08347*, 2024. 1

[16] 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