CVPR 2025 Roboflow-20VL Few-Shot Object Detection Challenge

Xuanlong Yu
Intellindust AI Lab
yu.xuanlong1996@gmail.com

Xi Shen
Intellindust AI Lab
shenxiluc@gmail.com

1. Strategy

We apply MMGrounding-DINO [5] and LLMDet [1] as our baseline models, and improve the few-shot ability via two branches: Few-shot fine-tuning and test-time augmentation. Based on the challenge dataset [3], all our models are trained on the provided training set without extra data. We choose the best models according to the validation set for the post-processing, while the mAP we report in this paper are the average results on the provided test sets.

1.1. Baseline Boosting

We choose MMGrounding-DINO with Swin-large backbone as our base model, and it has been pre-trained on O365V2, OpenImageV6, GoldG, V3det, COCO2017, LVISV1, COCO2014, GRIT, RefCOCO, RefCOCO+, RefCOCOg, and gRefCOCO.

Hyperparameter Ablation. We first tuned the training hyperparameters for the baseline MMGrounding-DINO model. A key finding was that the image and language backbones should be at least 10 times smaller than the detector; otherwise, the model readily overfits. Additionally, the varying number of annotations across datasets, especially with a differing number of classes, exacerbates overfitting on specific datasets, complicating model tuning. To mitigate this challenge and simplify the training procedure, we adopted the Plateau learning rate scheduler. As shown in Table 1, the Plateau scheduler and correct learning rates for backbones can improve the few-shot detection results.

Dataset-Specific Data Augmentation. We observed that for some of the datasets, we can not only apply traditional horizontal flipping (Hflip), but also possible to apply vertical flipping (Vflip) and diagonal flipping (Dflip). For instance, the datasets with aerial or bird's-eye views, as well as datasets with densely packed objects. Table 1 shows that, with multiple kinds of flipping, the performance can further improve. Moreover, this makes it possible to apply multiple flipping test-time augmentation (TTA), which shows strong potential for boosting the performance.

Base Model Adjustment. We notice that not only has MMGrounding-DINO fully open-sourced the large open-vocabulary object detection model, but recently, LLMDet [1] also provides a model checkpoint with the identical structure as MMGrounding-DINO with better performance on various benchmarks. LLMDet directly adapts MMGrounding-DINO as the base model and further improves it by tuning it on a new GroundingCap-1M dataset and uses LLaVA as the feature enhancer during training. Yet, we discovered that LLMDet was trained on the GroundingCap-1M dataset with freezing the Swin backbone, also, the MMGrounding-DINO baseline that LLMDet chose was pre-trained on O365V2, OpenImageV6, and GoldG, which was not the version with full datasets that were listed at the beginning of the section. We consider that, more images the image backbones learned, the better transfer ability the backbone has. Therefore, we adapt the weights of Bert and Detector in LLMDet to the MMGrounding-DINO base model we chose, to fully use the potential of the scaling of the pre-training. We denote this backbone as MMGDINO-LLMDet as shown in Table 1. We can see that the new base model can further improve the few-shot detection result, under the same data augmentation, scheduler, and test-time augmentation.

1.2. Test-Time Augmentation

Multiple Direction Flipping. Since we discovered the improvement given by flipping strategies, we also apply different directions of flipping during inference. Furthermore, to combine the results given by different flippings, we compare the performance between NMS and Soft-NMS. Table 2 illustrates the ablation results. The improvement is incremental and convincing.

Deep Ensembles using Weighted Box Fusion. Deep Ensembles [2] can effectively improve the detection performance by combining the results given by different models. After training several models according to strategies in Table 1, we obtain nine models with identical structure yet different weights. Different from the original Deep Ensembles, it is not trivial to directly average the bounding box

Base Model	**MMGDINO**					**MMGDINO-LLMDet**
Nb Epoch	12	12	24	24	24	24
Detector LR	2e-4	2e-4	2e-4	2e-4	2e-4	1e-4
Backbone LR	2e-5 Swin 2e-5 Bert	2e-4 Swin 2e-5 Bert	2e-5 Swin 2e-5 Bert	1e-5 Swin 1e-5 Bert	1e-5 Swin 1e-5 Bert	1e-5 Swin 1e-5 Bert
Scheduler	Milestone	Milestone	Plateau	Plateau	Plateau	Plateau
Flipping Aug	Hflip	Hflip	Hflip	Hflip	Vflip / Hflip / Dflip	Vflip / Hflip / Dflip
Flipping TTA	-	-	-	-	Hflip	Hflip
mAP	38.25	34.86	39.615	39.99	40.8	41.97
AP50	60.235	56.02	62.905	63.555	65.123	65.465

Table 1. **Ablation study on Baseline Boosting**

Base Model	**MMGDINO-LLMDet**		
Nb Epoch	24		
Detector LR	1e-4		
Backbone LR	1e-5 Swin 1e-5 Bert		
Scheduler	Plateau		
Flipping Aug	Vflip / Hflip / Dflip		
Flipping TTA	Hflip + NMS	Vflip / Hflip / Dflip + NMS	Vflip / Hflip / Dflip + soft-NMS
mAP	41.97	42.05	42.48
AP50	65.47	65.50	65.62

Table 2. **Ablation study on different flipping and NMS techniques during inference.**

scores in object detection. Thus, we use weighted box fusion [4] to fuse the model predictions, which greatly improves our results on the test set. We found that only two model ensembles can already improve mAP from 41.97 to 42.93. Yet we further noticed that the improvement becomes marginal when the number of models increases and the individual ability improves. For example, when it comes to five models ensembles, mAP increases to 44.729, while a good base model in the ensembles can achieve 42.48 mAP. Finally, we chose nine models for the final ensembles and achieved our final result mAP 46.187.

References

[1] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. *arXiv preprint arXiv:2501.18954*, 2025. 1

[2] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30, 2017. 1

[3] Anish Madan, Neehar Peri, Shu Kong, and Deva Ramanan. Revisiting few-shot object detection with vision-language models. *arXiv preprint arXiv:2312.14494*, 2023. 1

[4] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, pages 1–6, 2021. 2

[5] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 1