A Technical Report for CVPR 2026 Roboflow-20VL Few-Shot Object Detection Challenge

Chenhao Zhou[1,2] Qianqian Xu[3,4,*] Minye Lei[1,2] Yihang Huang[1,2]
Peisong Wen[2] Siran Dai[5,6] Yang Liu[2] Qingming Huang[2,3,*]
[1]Key Lab. of Intelligent Information Processing, Institute of Computing Technology, CAS
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences
[3]State Key Laboratory of AI Safety, Institute of Computing Technology, CAS
[4]Beijing Academy of Artificial Intelligence
[5]Institute of Information Engineering, CAS
[6]School of Cyber Security, University of Chinese Academy of Sciences
huangyihang0411@gmail.com

Abstract

Vision-language models (VLMs) such as GroundingDINO achieve strong zero-shot object detection on natural-image benchmarks, but their performance drops on Roboflow-20VL domains such as dental X-rays, aerial imagery, industrial inspection, and structured interface images. This report describes our solution for the In-Context Prompting Track of the Foundational Few-Shot Object Detection Challenge, where model fine-tuning is not allowed. We adapt frozen VLMs through domain-specific prompt design guided by few-shot examples. Our system combines single-pass multi-class prompting, class-specific prompting for frequently missed or confused categories, and a generative overlay paradigm for subsets with severe domain shifts. We further use an MLLM-as-a-Judge procedure to rescore candidate boxes according to category correctness and localization quality. Experiments on Roboflow-20VL show that careful prompt engineering can yield competitive few-shot detection performance without parameter updates.

1. Introduction

Vision-Language Models (VLMs) and open-vocabulary detectors such as GroundingDINO and GroundingDINO 1.5 [2, 4] have shown strong zero-shot object detection ability on standard natural-image benchmarks such as COCO [1]. However, their performance often degrades in the RF20VL setting [5], which contains highly diverse domains including dental X-rays, aerial imagery, industrial inspection images, and structured interface screenshots. These domains introduce substantial distribution shifts and semantic ambiguities, making category names alone insufficient for reliable grounding.

A common solution is gradient-based fine-tuning on the few available labeled examples [3]. This strategy requires model access and non-trivial computation, and is explicitly disallowed in the In-Context Prompting Track of the Foundational Few-Shot Object Detection Challenge. The central problem is therefore how to adapt a frozen VLM to domain-specific concepts using only limited labeled examples as references for prompt design.

We propose a compact prompt engineering framework for this setting. Instead of using one uniform prompt, we select domain-specific strategies according to observed failure modes. For most subsets, the model detects all target categories in one pass; for selected difficult categories, class-specific prompts reduce misses and inter-class confusion. For subsets with severe domain shifts, we use a multimodal generative model [7] to draw visual box overlays and recover proposals through deterministic image processing, avoiding unreliable direct coordinate prediction.

Finally, we introduce an MLLM-as-a-Judge confidence rescoring mechanism [9]. The judge inspects each rendered candidate box and assigns a score based on category correctness and localization quality, producing more useful confidence values than direct score generation. Combining tailored prompts, generative proposals, lightweight box refinement, and MLLM-based rescoring, our method achieves competitive performance on Roboflow-20VL without any parameter updates.

*Corresponding authors.

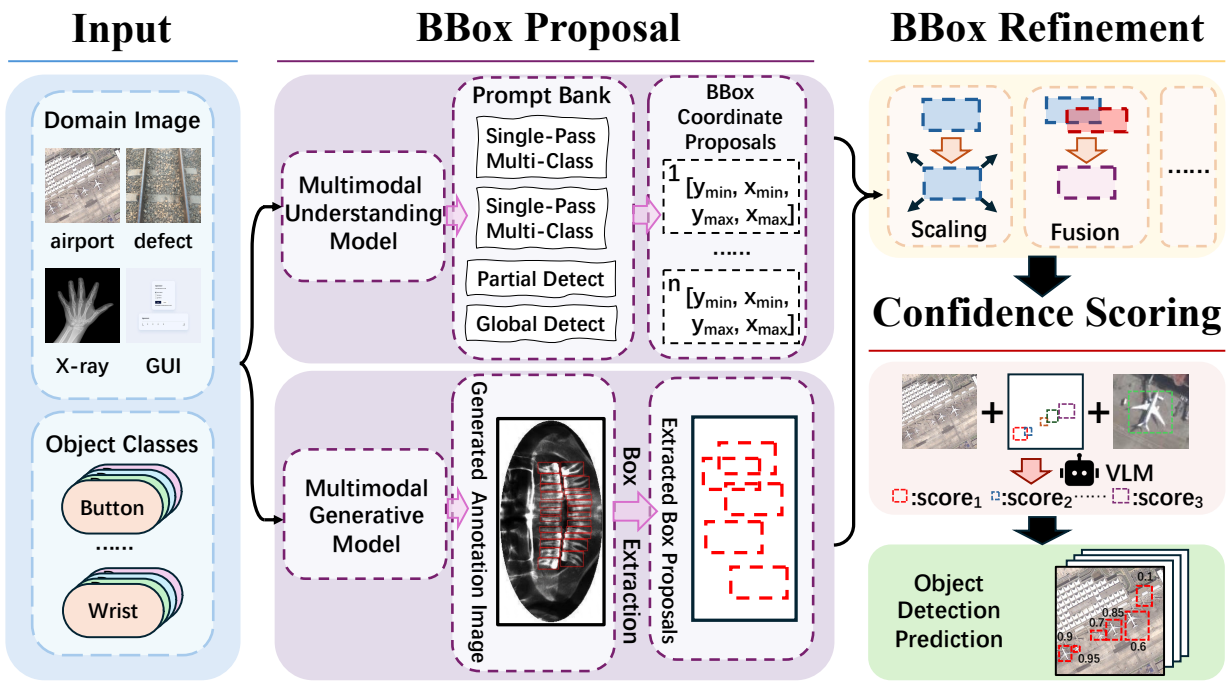

Figure 1. Overview of our framework.

2. Method

2.1. Bounding Box Proposal

2.1.1. Bounding Box Proposal with a Multimodal Understanding Model

We employ Gemini 3.5 Flash [7] as the multimodal understanding model for generating object bounding box proposals. The model is selected for three main reasons. First, it demonstrates strong vision-language grounding ability, enabling it to localize objects based on open-vocabulary textual descriptions without requiring task-specific detector training. Second, its instruction-following capability allows us to enforce a structured output format, including class labels and normalized bounding box coordinates. Third, the Flash variant provides a favorable trade-off between inference efficiency and multimodal reasoning quality, which is important when processing a large number of domain-specific images.

Given an input image and a set of target categories from a specific domain, the model is prompted to identify visible instances and return their bounding boxes in a predefined format. Each proposal consists of a class label and a bounding box represented by normalized coordinates. These generated boxes serve as initial object proposals for subsequent refinement.

Single-Pass-Multi-Class In the Single-Pass-Multi-Class setting, the model is prompted once with all candidate categories relevant to a given domain image. The model is asked to detect every visible object instance belonging to any of the provided categories and to output the corresponding class label and bounding box for each instance. This strategy allows the multimodal model to jointly reason over multiple semantic categories within the same image.

A key advantage of this formulation is that the model can compare and disambiguate visually similar categories in a shared context. Instead of treating each category independently, the model can leverage relative appearance, spatial layout, and contextual cues to decide which class each object belongs to. In practice, this joint prompting strategy improves both grounding quality and category assignment, especially when multiple target classes co-occur in the same image. It also reduces inference cost by requiring only one model call per image.

Single-Pass-Single-Class In the Single-Pass-Single-Class setting, the model is prompted with only one target category at a time. The prompt explicitly asks the model to focus on locating all visible instances of that specific category while ignoring objects from other classes. This class-specific formulation is particularly useful for categories that are frequently missed or confused in the multi-class setting.

Compared with the multi-class prompt, the single-class prompt reduces semantic competition among categories and encourages the model to allocate more attention to subtle or rare object instances. Therefore, we use this strategy for categories whose detection performance is insufficient under the Single-Pass-Multi-Class setting. For these selected categories, we issue additional class-specific prompts and directly replace the corresponding category predictions from the multi-class results with the single-class predictions. The final proposal set is then formed by combining the unchanged multi-class predictions for the remaining categories with the substituted single-class predictions for the difficult categories.

2.1.2. Bounding Box Proposal with a Multimodal Generative Model

While the multimodal understanding model works well for many subsets through single-pass multi-class or single-pass single-class prompting, we observe that this paradigm becomes less reliable under severe domain shifts. In highly specialized images, the visual appearance of objects may differ substantially from natural-image concepts, and category names alone may not provide sufficient semantic grounding. Therefore, for the most challenging subsets, we introduce a multimodal generative model-based bounding box proposal strategy [7] as a training-free alternative.

Instead of asking the model to directly predict bounding box coordinates, we prompt an image generation model to draw class-specific colored rectangle outlines on the original image. The prompt describes the domain-specific visual cues of each target category and instructs the model to preserve the original image content, resolution, brightness, and aspect ratio, while only adding bounding box overlays.

This formulation converts object localization into a visual annotation task. The generative model is encouraged to mark regions according to visual patterns specified in the prompt, rather than relying solely on open-vocabulary detection ability. The generated colored boxes are then recovered by deterministic image processing, which makes the final proposal extraction independent of textual coordinate generation.

Specifically, each target category is assigned a unique color. After the model returns the annotated image, we convert it into HSV color space and apply color-specific thresholding to isolate the box outlines for each category. The extracted masks are further processed by contour detection or rectilinear line recovery to obtain bounding rectangles.

Finally, the recovered boxes are mapped back to the original image coordinate system. If the generated image size differs from the input image size, the coordinates are rescaled accordingly. The resulting proposals are saved in COCO-style format [1] as

$$[x, y, w, h],$$

with their corresponding category labels.

2.2. Bounding Box Refinement

After obtaining the initial bounding box proposals, we further refine their spatial extent by comparing model predictions with ground-truth annotations on the training and validation sets. We observe that many errors are not caused by incorrect localization, but by systematic differences between the model's visual grounding behavior and human annotation preferences. In other words, the model often identifies the correct object region, while the predicted box extent may be consistently smaller, larger, or biased toward a particular object part. To address this issue without model training, we apply lightweight, class-specific bounding box refinement strategies.

Weighted Box Fusion. For categories whose annotation preference lies between local and global object regions, we use Weighted Box Fusion (WBF) [6] to combine multiple prompt-induced box proposals. Specifically, we design different prompts to guide the model toward different spatial interpretations of the same object. For example, for soda bottles, one prompt may encourage the model to localize a discriminative local part such as the bottle cap, while another prompt asks the model to annotate a more complete visible region including both the cap and the bottle body. These different predictions provide complementary spatial cues.

Since confidence scores are not yet available at this stage, we assign predefined fusion weights to different prompting strategies and treat them as tunable hyperparameters. Given multiple boxes predicted for the same object, WBF merges them into a single refined box by computing a weighted average of their coordinates:

$$B_{\text{fused}} = \frac{\sum_i \lambda_i B_i}{\sum_i \lambda_i},$$

where B_i denotes the box produced by the i-th prompting strategy, and λ_i is its corresponding fusion weight. The weights are selected based on the alignment between predictions and human annotations on the labeled data. This fusion strategy allows the final box to better match human annotation preferences than either local-only or global-only prompting alone.

Class-Specific Box Scaling. For some categories, we observe a consistent scale bias between predicted boxes and ground-truth annotations. For example, in specialized domains such as X-ray images, the model can correctly localize the target anatomical structure, but the predicted box may cover a slightly smaller region than the expert annotation. In such cases, we apply class-specific box scaling

to adjust the predicted extent while keeping the box center unchanged.

Given a predicted box $B = (x, y, w, h)$, we compute its center (c_x, c_y) and rescale its width and height by class-specific factors α_w and α_h:

$$w' = \alpha_w w, \quad h' = \alpha_h h.$$

The refined box is then reconstructed as

$$B' = \left(c_x - \frac{w'}{2},\ c_y - \frac{h'}{2},\ w',\ h' \right).$$

The scaling factors are selected according to the systematic bias observed on the labeled data. This simple refinement is especially effective when the model's localization is reliable but its box extent is misaligned with the annotation protocol.

2.3. Confidence Rescoring with Multimodal Understanding Model

The initial bounding box proposal stage mainly focuses on localization, while the confidence scores produced together with the boxes are often poorly calibrated. In particular, when the model is directly asked to output both bounding boxes and confidence values, it tends to assign conservative and weakly discriminative scores, such as values around 0.6. Such scores are insufficient for ranking predictions and suppressing false positives. To address this issue, we introduce an additional confidence rescoring stage based on a multimodal understanding model, inspired by the MLLM-as-a-Judge paradigm [9].

Instead of asking the model to judge a box only from its numerical coordinates, we reformulate confidence estimation as a visual verification task. For each predicted instance, we render the candidate bounding box on the original image using a red rectangle. The multimodal model is then asked to inspect the image and determine whether the red box correctly covers an object of the target class. This design leverages the model's visual reasoning ability directly, rather than relying on its ability to calibrate confidence during coordinate generation.

Optionally, we provide an in-context visual exemplar from the training set. For each category, we select a representative annotated training instance and draw its ground-truth box in green. The exemplar image is given together with the test image, allowing the model to compare the candidate box against both the target appearance and the human annotation preference for box tightness. Thus, the input to the rescoring model consists of either one image, i.e., the test image with a red candidate box, or two images, i.e., a green-box training exemplar followed by the red-box test candidate.

The prompt specifies the dataset name, target class, candidate box in COCO pixel format [1], and image size. The model is instructed to return a single JSON object:

$$\{\texttt{"score"} : s\},$$

where $s \in [0, 1]$. The scoring rubric encourages high scores for correct and well-localized detections, medium scores for shifted, loose, or partial boxes, and low scores for wrong-class predictions, weak overlaps, or false positives. In our implementation, the model is queried with deterministic decoding and JSON-formatted output to ensure stable parsing.

Finally, the parsed score is clipped to $[0, 1]$ and used to replace the original confidence score of the corresponding prediction. This rescoring step does not modify the predicted class label or bounding box coordinates. It only updates the ranking confidence of each proposal, enabling false positives and poorly localized boxes to be suppressed during evaluation.

3. Results

Table 1. Ablation study of single-pass-single-class (for recode-waste dataset)

Method	mAP	Δ
Multi-class	33.97	-
Multi-class + Single-class	**36.19**	+2.22

Table 2. Ablation study of generative method (for X-ray-id dataset)

Method	mAP	Δ
Understanding-based	1.51	-
Generative-based	**21.23**	+19.72

Table 3. Ablation study of confidence rescoring (overall)

Method	mAP	Δ
MM-GroundingDINO Swin-L [8] (Zero-shot)	16.79	–
w/o confidence	33.38	+16.59
w/ confidence	**35.67**	+18.88

Tables 1–3 summarize the effectiveness of the main components in our pipeline. For the recode-waste subset, the aggregate category is frequently missed or confused under the single-pass-multi-class setting. We therefore apply single-pass-single-class prompting to this category and merge the resulting aggregate detections with the original multi-class predictions. This simple replacement improves mAP from 33.97 to 36.19, showing that class-specific prompting can

complement joint multi-class reasoning for difficult categories. For the X-ray-id subset, the generative model-based proposal strategy improves over the multimodal understanding baseline, indicating that converting localization into a visual box-overlay task is more robust under severe domain shift. Finally, MLLM-based confidence rescoring improves the overall mAP from 33.38 to 35.67.

4. Conclusion

We present a training-free solution for the In-Context Prompting Track of the CVPR 2026 Roboflow-20VL Few-Shot Object Detection Challenge. Our framework adapts frozen vision-language models through domain-specific prompting, combining multi-class prompting, class-specific prompting for difficult categories, generative box-overlay proposals for severe domain shifts, and MLLM-based confidence rescoring. Experimental results show that these simple components consistently improve detection performance, demonstrating that careful prompt design and visual verification can effectively enhance few-shot object detection without parameter updates.

References

[1] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 1, 3, 4

[2] 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. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, pages 38–55. Springer, 2024. 1

[3] Anish Madan, Neehar Peri, Shu Kong, and Deva Ramanan. Revisiting few-shot object detection with vision-language models. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 1

[4] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding DINO 1.5: Advance the "edge" of open-set object detection. *CoRR*, abs/2405.10300, 2024. 1

[5] 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. *CoRR*, abs/2505.20612, 2025. 1

[6] Roman A. Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image Vis. Comput.*, 107:104117, 2021. 3

[7] Gemini Team. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023. 1, 2, 3

[8] 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. 4

[9] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1, 4