Discover the Unknown and Reconsider the Known: Specializing Multimodal Promptable Detectors for Diverse Domains

Kyeongryeol Go* Hyundong Jin* Taewoong Jang Wooseong Choi

Superb AI
Seoul, South Korea

{krgo, hyundong.jin, twjang, wschoi}@superb-ai.com

Abstract

Existing open-set vision foundation models (VFMs) rely predominantly on text prompts to specify target objects. However, they often suffer from semantic ambiguity in specialized domains with ambiguous category names. To alleviate this ambiguity, we employ ZERO, a pretrained VFM that grounds target objects from multimodal prompts (text and visual cues), allowing a visual cue to complement an ambiguous category name. In addition, to better localize the target object, ZERO leverages the surrounding scene of each prompt, enabling robust and context-aware grounding. The Foundational Few-Shot Object Detection (FSOD) challenge comprises highly specialized, sparsely annotated domains. To adapt ZERO to these domains, we first discover informative prompts by searching attribute-grounded descriptions and additional labels for enriching its sparse annotations. After adaptation, we reconsider the resulting detections to correct their labels for alleviating spurious detections. Experimental results show that our approach achieves an mAP of 53.86 averaged across the 20 domains, demonstrating its effectiveness despite their heterogeneity.

1. Introduction

Foundation models, large-scale neural networks on a large amount of data, have become a central paradigm in modern AI [2, 3, 7, 14]. Through such pretraining, these models acquire rich representations that generalize across diverse tasks. In practice, rather than training a dedicated model from scratch, a pretrained model is adapted to each new task, significantly reducing the need for large task-specific annotated datasets.

Such generalization is also required for tasks that detect objects within a given image. Conventional detectors [4, 11, 15] are trained to recognize a fixed set of predefined categories (closed-set), which limits their applicability to novel or evolving domains. In contrast, vision foundation models (VFMs) [10, 12, 18] recognize and localize arbitrary objects beyond a predefined set. This open-set recognition is valuable in real-world deployments, where the objects of interest are diverse or often unknown at training time. By allowing target objects to be specified through prompts, VFMs can detect novel objects, reducing the need for retraining.

Despite these advancements, existing open-set VFMs are predominantly driven by text prompts, which specify the target object by its category name in natural language (e.g., GLIP [10] and Grounding DINO [12]). Detecting objects queried by text is intuitive and effective for common categories, but it becomes unreliable in unfamiliar domains—such as manufacturing, medical, or surveillance—where category names are semantically ambiguous or underspecified [9]. An alternative is querying with a visual prompt, which specifies the target with an example image, thereby avoiding such semantic ambiguity. However, a visual prompt typically provides an exemplar of the object without its surrounding context, which limits the available cues for identifying the target object [8, 19].

To address these limitations, we employ ZERO [6], a vision foundation model from Superb AI, promptable with multimodal cues for open-set visual grounding. A target object can be specified through a text prompt, a visual prompt, or both, allowing the modalities to reinforce each other against ambiguous specifications. Moreover, ZERO grounds each visual prompt together with its surrounding scene, exploiting contextual cues for more robust localization. Pretrained on a high-quality dataset curated from both general and industrial sources, ZERO achieves strong zero-shot performance across diverse domains. We further adapt ZERO to highly specialized, sparsely annotated domains under the Foundational Few-Shot Object Detection (FSOD) challenge [13, 16]. We first discover informative prompts and additional labels to fine-tune the model. Then we re-

*Equal contribution.

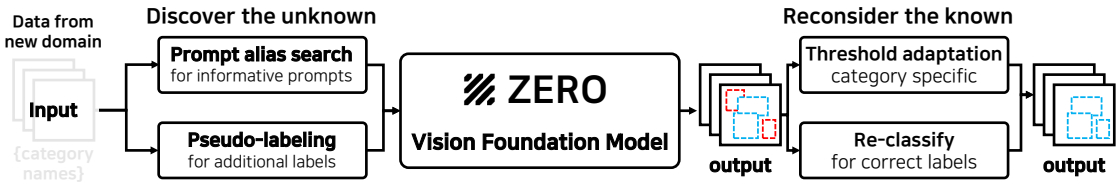

Figure 1. For each domain, *Discover the unknown* obtains informative prompts through prompt alias search and additional labels through pseudo-labeling. After discovering the unknown, we fine-tune our vision foundation model, ZERO, to the target domain. *Reconsider the known* then refines the resulting detections with category-specific threshold adaptation and re-classification, correcting their labels and removing spurious boxes. Correct detections are shown in blue, and spurious or mislabeled ones in red.

consider its proposed localized regions to correct their predicted categories and remove spurious predictions. Across 20 visually diverse domains, ZERO achieves strong performance, demonstrating that it offers a practical foundation for object detection in specialized real-world domains. An overview of the proposed framework is presented in Figure 1.

2. Methodology

2.1. Discover the unknown

We aim to fine-tune ZERO on the few-shot data of each target domain. However, two obstacles hinder effective adaptation: the category names provided in the datasets may not offer informative grounding queries, and the annotations are sparse. We discover the prompts and labels before the fine-tuning stage.

Prompt alias search. Under sparse annotation, a text prompt offers a label-efficient specification of each target category, requiring no additional annotation. However, detection performance is dependent on the extent to which this prompt characterizes the visual appearance of the target object. This dependence becomes particularly acute in the FSOD challenge, where each category prompt defaults to the class name provided in the dataset [16], which is frequently uninformative in specialized domains. An abbreviated class name, for example, provides limited visual information and constitutes a weak grounding query.

To address this, we augment each category with a set of alias prompts: paraphrases that describe its visual attributes rather than its name. Such prompts strengthen the grounding signal, but may not be beneficial for all domains. Some aliases drift toward neighboring concepts or introduce false positives, and their usefulness varies across domains; adding all of them indiscriminately can degrade accuracy. We therefore select aliases greedily and per domain: starting from the class names, we iteratively add the alias that most improves detection accuracy and stop when no further gain remains. The search remains tractable because a single cached forward pass of ZERO suffices to score any alias subset without re-running inference. The selected prompts are then used in fine-tuning and post-processing.

Pseudo-labeling. The annotations provided for each domain are sparse, leaving many object instances in the images unlabeled. Fine-tuning ZERO directly on such data would treat the unlabeled objects as background, teaching the model to suppress valid detections. We therefore require additional annotations to cover the missing instances. However, manual annotation is costly and often demands domain expertise.

To obtain these annotations without manual effort, we generate pseudo annotations, each comprising a bounding box and a category label. To this end, we employ our pretrained detector ZERO with off-the-shelf models, SAM3 [5] and Qwen3-VL-32B [20]. Given a query image and a text prompt specifying the target object, ZERO and SAM3 produce bounding boxes that enclose the corresponding objects, each with a confidence score. We retain the boxes whose score exceeds a predefined threshold. For each retained box, we crop the enclosed region and ask Qwen3-VL-32B to classify it into the most visually similar category defined for the target domain, or into an auxiliary "*unknown*" category whose crops are discarded as false positives. Finally, the resulting pseudo annotations are combined with the few ground-truth annotations to form an enlarged training set on which ZERO is adapted to the target domain.

2.2. Fine-tuning ZERO

To adapt ZERO effectively to each domain, we must set several hyperparameters that interact with one another: the learning rate, the strength of paraphrase augmentation, and the choice of trainable modules. As naïve approaches, one might consider an exhaustive per-domain search over the joint hyperparameter space, or a single configuration shared across all domains. However, the exhaustive search is computationally infeasible, while the shared configuration is suboptimal, as the domains vary widely in appearance.

We therefore decompose the search into two sequential stages. In the first stage, we select the learning rate and the paraphrase augmentation strategy (none, partial, or full). In the second stage, we fix the hyperparameters selected in the first stage, and determine which part of the model to update: the detection backbone, the detection heads, or the

language adapters. We adopt this staged search based on the empirical observation that the optimal learning rate and augmentation strategy remain largely stable regardless of which modules are trained. We run this search independently for each domain, retaining its best configuration rather than a single global one. Note that this search remains computationally affordable given the small amount of data in each domain. At inference, each domain is evaluated using the model trained under its selected configuration.

2.3. Reconsider the known

A detector fine-tuned on a few labels reliably learns where objects are but often mistakes what they are. Because these misclassifications are not consistently less confident than correct detections, a higher threshold cannot remove them without discarding correct detections as well.

We address this with a per-domain post-processing step of two components that can be applied independently. The first component refines detection: we reuse the enriched prompts from alias search and assign a separate confidence threshold to each category. The second component reconsiders the category assigned to each confident detection. We train a lightweight classifier on the few-shot crops and apply it to detections that exceed their per-category threshold; we leave low-confidence detections untouched, as few-shot training leaves the classifier unreliable there and its predictions would corrupt the labels. The classifier predicts a category for each crop together with a probability, which we treat as its confidence. When this confidence exceeds an acceptance threshold, we keep the existing label if the prediction matches it and reassign the category otherwise; when it stays below, we remove the detection. In effect, the detector acts as a high-recall proposal generator, the prompts and thresholds control where objects are detected, and the classifier reconsiders the label assigned to each.

2.4. Test-time augmentation

At inference, we strengthen the detections of each adapted model with test-time augmentation (TTA) applied entirely on the client side: every transformation is performed on the input image before the detector is queried, and the returned boxes are mapped back into the original-image coordinates, so the detector interface remains a plain promptable query.

We employ three complementary augmentations. *Multi-scale* inference queries the detector at several shortest-edge resolutions, each capped to preserve the detector's default longest-edge aspect ratio, which helps recover objects whose scale departs from the pretraining distribution. *Horizontal flipping* mirrors the image and reflects the predicted boxes back, providing an independent view of each scene. *Tiling* partitions large images into overlapping crops and detects within each, recovering small or densely packed objects that are easily missed at full resolution; the per-tile detections are translated to global coordinates and clipped to the image bounds. The detections from all augmented views are then pooled per prompt and de-duplicated with class-aware non-maximum suppression followed by top-k selection.

2.5. Multi-source fusion

The proposed framework yields several complementary detection sources for each domain: models adapted under different configurations, different backbone scales, and the text and visual prompting modalities. We combine them with a train-free, detector-free fusion stage that operates purely on the saved detections, requiring no further inference.

We first exploit the structure of the evaluation metric. Because COCO mAP is the mean of per-category average precision and each category's score is independent of the others, we route categories independently: for every category we keep the detections from the source that scores best on it, and concatenate the results across categories. This per-category selection is strictly finer-grained than choosing a single best source per domain, and can only match or improve it on the tuning metric.

When several sources are individually strong for a category, selection alone discards complementary boxes. We therefore additionally fuse the top-ranked sources with Weighted Boxes Fusion [17], which clusters overlapping boxes across sources and replaces each cluster with a confidence-weighted average box, with sources weighted either equally or in proportion to their validation accuracy. We also consider Soft-NMS [1] and a rank-decayed NMS merge as alternative fusion operators. For each domain, we search over these operators and their hyperparameters and retain the best-scoring recipe.

All source selection and fusion search is scored on the validation split; the chosen recipe is then frozen and applied unchanged to the held-out test split, so that no test annotation influences the configuration.

3. Experiments

Evaluation protocol. We evaluate ZERO, adapted to each target domain through the proposed framework, on Roboflow20-VL, the 20-dataset subset of Roboflow100-VL [16] used in the Foundational FSOD challenge, where each category is specified by 10 annotated examples (10-shot). Following the challenge, each domain is evaluated independently using the COCO-style average precision, averaged over IoU thresholds from 0.50 to 0.95. We report the mean average precision (mAP) averaged across the 20 domains as our primary metric.

Results. ZERO achieves an mAP of 53.86 across the 20 domains of Roboflow20-VL, performing reliably despite their wide variation in appearance.

4. Conclusion

We introduced ZERO, a pretrained open-set vision foundation model developed at Superb AI that grounds objects from text prompts, visual prompts, or their combination, using the surrounding scene as context. Trained on a high-quality dataset curated from general and industrial sources, ZERO provides strong representations that generalize across a wide range of domains in the zero-shot setting. To adapt ZERO to the specialized, sparsely annotated domains of the Foundational FSOD challenge, we discover informative prompts and additional labels, then reconsider the resulting detections to correct their categories and remove spurious ones. Experimental results on Roboflow20-VL show that ZERO achieves an mAP of 53.86 across 20 visually diverse domains, indicating its strong adaptability to specialized real-world settings.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5561–5569, 2017. 3

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1

[5] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025. 2

[6] Sangbum Choi, Kyeongryeol Go, and Taewoong Jang. ZERO: Industry-ready vision foundation model with multimodal prompts. *arXiv preprint arXiv:2507.04270*, 2025. 1

[7] 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*, pages 4171–4186, 2019. 1

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1

[9] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-Rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pages 38–57. Springer, 2024. 1

[10] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for ense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 1

[12] 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. 1

[13] 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

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021. 1

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016. 1

[16] 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. *Advances in Neural Information Processing Systems*, 2025. 1, 2, 3

[17] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107: 104117, 2021. 3

[18] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YoloE: Real-time seeing anything. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 24591–24602, 2025. 1

[19] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7031–7040, 2023. 1

[20] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2