ZERO : Multi-modal Prompt based Visual Grounding

Sangbum Choi
SuperbAI
Seoul, South Korea
sbchoi@superb-ai.com

Kyeongryeol Go
SuperbAI
Seoul, South Korea
krgo@superb-ai.com

Abstract

Recent advances in artificial intelligence have led to the emergence of foundation models, large-scale pre-trained neural networks that serve as versatile starting points for a wide range of downstream tasks. In this work, we present ZERO, a zero-shot multi-prompt object detection model specifically designed for robust, production-ready deployment across diverse industrial domains. ZERO integrates direct image input with multiple user-defined prompts, which can include both textual and visual cues, and processes them through dedicated encoders to generate accurate detection outputs. The model architecture is optimized for scalability, with a total of 1.033 TFLOPS and 622.346 million parameters, and is trained using a domain-specific image database exceeding one billion images. For the CVPR 2025 Foundational Few-Shot Object Detection (FSOD) Challenge, we introduce a domain-specific fine-tuning strategy that emphasizes prompt diversity and conservative pseudo-labeling, enabling effective adaptation to new domains with minimal supervision. Our approach demonstrates practical advantages in flexibility, efficiency, and real-world applicability, achieving strong performance on the RF20VL-fsod benchmark despite limited annotation budgets. The results highlight the potential of prompt-driven, data-centric AI for scalable and adaptive object detection in dynamic industrial environments.

1. Introduction

Recent advances in artificial intelligence have been marked by the emergence of Foundation Models (FMs), which differ fundamentally from traditional, well-defined tasks such as object detection, text classification, instance segmentation, and fraud detection. Unlike these conventional tasks, which have clear objectives and boundaries, foundation models are not easily confined to a single, explicit definition. As highlighted in Rishi's report [15], foundation models encompass large-scale, pre-trained models—such as BERT [6], DALL-E [1], and GPT-3 [3] that are trained on vast datasets and can be adapted to a wide variety of downstream applications. Major technology companies similarly define foundation models as large deep learning neural networks trained on massive datasets, fundamentally changing how data scientists approach machine learning. Rather than building AI systems from scratch, practitioners now use foundation models as starting points, enabling faster and more cost-effective development of new applications.

In the vision domain, Vision Foundation Models (VFMs) are distinguished by their capacity to learn from diverse domains and support transfer learning, while also being capable of performing a variety of functions without further domain-specific training. Recent research has focused on models capable of detecting and segmenting arbitrary objects based on textual, visual, or prompt-based cues. The field has evolved from closed-set detection, where all classes are known during training, to open-set detection, which aims to identify both known and novel classes in dynamic, real-world environments. Many open-set models leverage pre-trained vision-language models like CLIP [13] to generalize to new concepts, but these often face limitations in aligning fine-grained image regions with textual descriptions. Approaches such as GLIP [8] and GroundingDINO [9] have introduced new architectures to address these challenges, supporting a broader range of prompts and tasks, though some limitations remain in terms of prompt flexibility and objective evaluation.

Our model named ZERO is developed with the goal of providing immediately deployable visual recognition capabilities in production environments, relying solely on prompting techniques. By leveraging a vast, domain-specific image database and a prompt-centric approach, it simplifies traditional data labeling and model training processes, allowing for rapid adaptation to new requirements and seamless integration with existing systems. This approach offers significant advantages in scalability, flexibility, and real-world applicability, supporting a wide array of domains and evolving business needs.

Model Component	FLOPS (G)
Image Backbone	547
Text & Visual Encoder	22
Transformer Layer	426
etc	28

Table 1. FLOPS for model architecture.

2. ZERO

2.1. Model Architecture

The internally developed model for object detection integrates direct image input (S) with multiple user-defined prompts $(P_i, i = 1, \ldots, N)$, which can range from generic keywords to simple sentences and are processed via a text encoder, while visual elements within prompts are handled by an image encoder, with both encodings fused to generate the final detection output. As shown in Figure 1, the transformer encoder fuses each independent representative feature. Individual FLOPS for each layer are shown in Table 1.

2.2. Training

When running the model on an NVIDIA A100-SXM4-80GB GPU with a mini-batch size of 2, approximately 65GB of GPU memory is consumed. Currently, the project is being conducted across 8 GPUs with a gradient accumulation step of 2, resulting in a total effective batch size of 32. For a dataset size of about 1M images, the training process takes approximately 240 hours.

2.3. Dataset

SuperbAI is a company that provides AI solutions for enterprises, specializing in rapidly building high-quality AI training data through an AI MLOps platform tailored for large-scale data management. The company's greatest strength is its possession of a vast domain-specific image database consisting of over 1 billion images. We have independently built industrial datasets that are available for use in current research. These datasets are highly relevant to industries such as manufacturing, distribution, logistics, security, and surveillance. From this pool, we have secured a dataset of 0.92M images available for current research and have defined 37 domains (including manufacturing, healthcare, security, autonomous driving, etc.), also preparing high-quality evaluation datasets for this purpose. In Figure 2, we described the brief pipeline for preparing our dataset. We used SAM2 [14] and captioning models such as smolVLM [11] and Qwen-VL [2] for generating captions.

3. Foundation FSOD challenge

CVPR 2025 Foundational FSOD Challenge [10, 16] focuses on performing few-shot object detection (FSOD) using the RF20VL-fsod dataset, curated from 20 distinct domains. Participants are allowed to pretrain their models on any external datasets, but fine-tuning must be conducted exclusively on the RF20VL-fsod dataset. The training split provides only 10 bounding box annotations per category, and due to this limited annotation budget, not all categories may be labeled in every image. Each category is accompanied by a noun phrase description that clarifies the meaning of category names, particularly those that are otherwise ambiguous or insufficiently descriptive. For example, the category name 'DIP' in the x-ray-id domain is clarified as 'Distal Interphalangeal joint, the farthest joint in the fingers'. The final evaluation score is computed by averaging the mAP@50:95 across all domains.

We propose a domain-specific fine-tuning of Superb AI's ZERO, a zero-shot multi-prompt object detection model that is robust enough for product-level deployment across various industrial domains. Leveraging our expertise in data-centric AI development, we designed our training pipeline to maximize performance under limited supervision. In the fine-tuning phase, we focus on promoting prompt diversity - both in textual and visual prompts - to enable effective adaptation to the heterogeneous and domain-specific nature of the RF20vl-fsod dataset. In this section, we explain the strategies used to adapt the model effectively to the target domains using limited annotations. It is structured into three stages: training, inference, and submission.

3.1. Training

During training, we introduce extensive prompt diversity to mitigate overfitting. At the text level, we employ LLaMA-3-8B-Instruct [5] to paraphrase category descriptions into concise noun phrases. The paraphrasing process is guided by a carefully crafted prompt template that instructs the model to preserve semantic integrity, avoid redundancy, and generate unambiguous and distinct definitions. This augmentation not only increases the lexical variety of prompts but also improves the model's ability to distinguish fine-grained categories. In addition to positive prompts, we include negative textual prompts to encourage a better discriminative embedding space obtained by contrastive learning. Note that the number of negative prompts is set with caution to avoid introducing label noise.

At the visual level, we apply both in-image and out-image prompting strategies. In-image visual prompts use objects co-occurring in the same image, reflecting a convention common in large-scale training in the vision foundation model [7] and aiding contextual reasoning. Out-image visual prompts, by contrast, introduce objects of the same category from other images, fostering generalization and re-

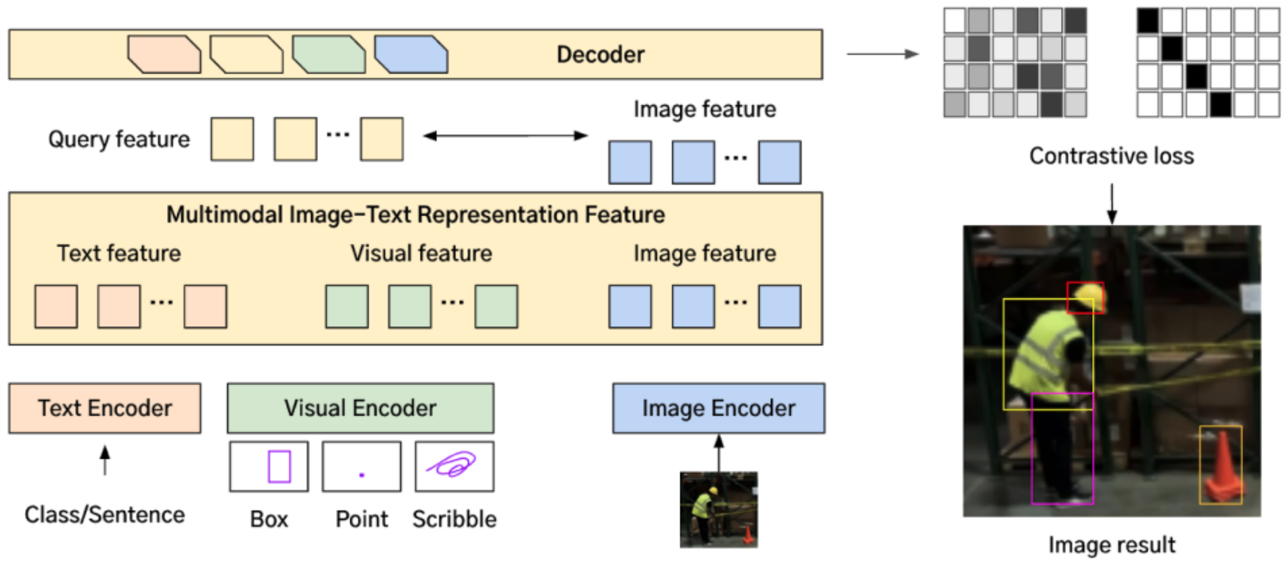

Figure 1. Model Architecture

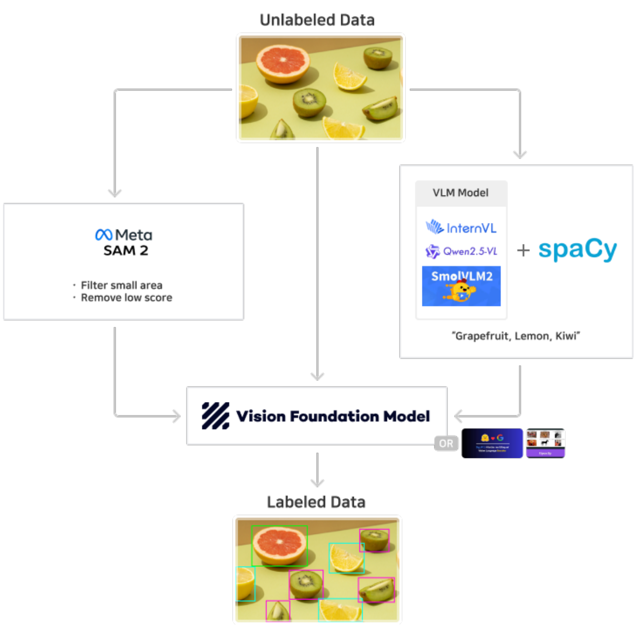

Figure 2. Data labeling pipeline

ducing reliance on narrow contextual cues.

Alongside prompt engineering, we implement a conservative pseudo-labeling strategy: only model predictions with high confidence scores are added as pseudo-labels for instances of unlabeled categories in the train split. This strict filtering ensures the integrity of additional supervision while preventing error accumulation from low-quality labels.

Instruction
You are an assistant specialized in generating concise noun-phrase definitions by paraphrasing. You will be given a list of terms in the format `[term] = [definition]`. For each term, return a corresponding line in the format `[term] = [paraphrased definition]`. Your paraphrased definitions must: 1. Be concise and written as noun phrases. 2. Preserve the original meaning and context. 3. Clearly distinguish each term from the others. 4. Follow the same line-by-line format as the input. Do not add or omit any terms.

Table 2. Instruction used for text prompt augmentation.

3.2. Inference

During inference, we adopt several techniques to improve detection performance and stability. Test-time augmentation (TTA) is applied by default using a combination of image resizing at multiple scales (0.8×, 1.0×, 1.2×) and horizontal flipping, effectively exposing the model to various spatial configurations. We also perform a category-wise threshold search, tuning the confidence threshold for each class individually to optimize the mean average precision (mAP) in the validation split. To further enhance robustness, we considered the ensemble of predictions from both text and visual prompts.

Factor	Options
Text prompt	`original` `original + augmented`
Visual prompt	`in-image` `out-image`
Annotations	`original` `original + pseudo-labeled`

Inference	`text`
	`visual`
	`text + visual`

Table 3. Factors considered for checkpoint selection.

3.3. Submission

Due to the distinctiveness of each domain, no universal strategy consistently outperforms others. Therefore, we obtained various model checkpoints by turning on and off the proposals for training and inference, and selected the best checkpoint guided by the performance on a validation split. For each dataset from the 20 distinct domains, the best checkpoint is selected based on combinations of four factors in Table 3.

It is important to note that, similar to the training split, the validation split is only partially annotated. Therefore, during checkpoint evaluation, we restrict predictions to the categories known to be present in each image. In contrast, for test split inference intended for final submission, predictions are obtained across all categories.

4. Conclusion

In this work, we presented a domain-specific fine-tuning strategy for adapting a zero-shot multi-prompt object detection model, Superb AI's ZERO, to the RF20vl-fsod few-shot detection challenge. By leveraging data-centric design principles, including prompt diversity and conservative pseudo-labeling, our approach effectively adapts to diverse domains using minimal supervision. Through carefully crafted textual and visual prompts, combined with robust inference techniques, we demonstrated a practical and scalable solution for domain-adaptive FSOD under limited annotation conditions.

As future work, we are exploring several directions to enhance performance and generalization further. These include selective fine-tuning techniques, such as BOIL [12], to balance generalization and specificity in few-shot learning, and adopting particular design choices, like NaViT [4], to better preserve aspect ratios and incorporate visual context during training and inference for improved spatial fidelity.

References

[1] Gabriel Goh Scott Gray Chelsea Voss Alec Radford Mark Chen Ilya Sutskever Aditya Ramesh, Mikhail Pavlov. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 1

[4] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023. 4

[5] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[6] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 1

[7] 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. 2

[8] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 1

[9] 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, page 38–55, Berlin, Heidelberg, 2024. Springer-Verlag. 1

[10] 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

[11] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 2

[12] Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. Boil: Towards representation change for few-shot learning. *arXiv preprint arXiv:2008.08882*, 2020. 4

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[14] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[15] Ehsan Adeli Russ Altman Simran Arora Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji Annie Chen Kathleen Creel Jared Quincy Davis Dora Demszky Chris Donahue Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou Percy Liang Rishi Bommasani, Drew A. Hudson. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2022. 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. *arXiv preprint arXiv:2505.20612*, 2025. 2