The Solution for Foundational Few-Shot Object Detection Challenge 2024

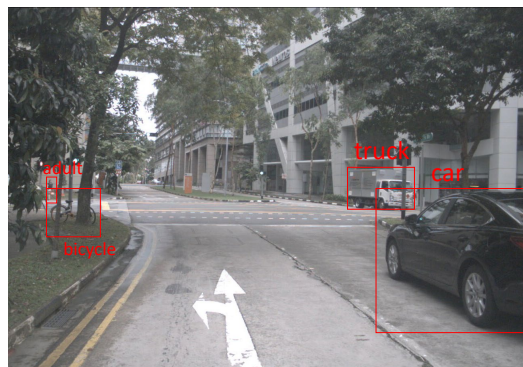Hongpeng Pan[1], Jinkai Li[1], Shifeng Yi[1], Bing Hu[2], Shouwei Yang[1], Yi Xu[2], Yang Yang[1,*]

[1]Nanjing University of Science and Technology [2]Dalian University of Technology

Abstract

This report introduces an enhanced method for the Foundational Few-Shot Object Detection (FSOD) task, leveraging the vision-language model (VLM) for object detection. However, on specific datasets, VLM may encounter the problem where the detected targets are misaligned with the target concepts of interest. This misalignment hinders the zero-shot performance of VLM and the application of fine-tuning methods based on pseudo-labels. To address this issue, we propose the VLM+ framework, which integrates the multimodal large language model (MM-LLM). Specifically, we use MM-LLM to generate a series of referential expressions for each category. Based on the VLM predictions and the given annotations, we select the best referential expression for each category by matching the maximum IoU in training set. Subsequently, we use these referential expressions to generate pseudo-labels for all images in the training set and then combine them with the original labeled data to fine-tune the VLM. Additionally, we employ iterative pseudo-label generation and optimization to further enhance the performance of the VLM. Our approach achieve 32.56 mAP in the final test.

1. Introduction

Deep learning techniques have garnered widespread attention across multiple research fields. [7, 9, 6, 8]. Object detection, as a fundamental task in computer vision, has garnered extensive research. Traditional visual recognition models are typically trained to predict a fixed set of predefined object categories, which limits their usability in real-world applications, as additional labeled data is required to generalize to new visual concepts and domains. To address this issue, some open-set object detection methods have been proposed, such as GLIP [3] and Grounding DINO [4]. These methods reframe object detection as a phrase-based task and introduce contrastive training between object re-

---

*Corresponding Author

Prompt: adult, bicycle, truck, car

Figure 1: By specifying the interested classes in textual prompts, VLMs can implement zero-shot object detection.
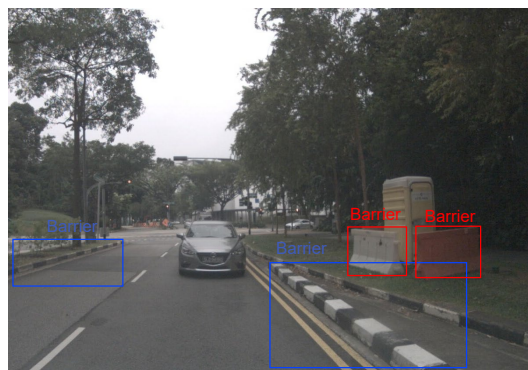

Figure 2: Poor Alignment Between VLM and Class Prompts. In the nuImages dataset, barriers are defined as road barricades (in red), while the obstacles predicted by the VLMs include roadside steps (in blue).

gions and language phrases. Due to their excellent alignment between textual and visual features, these models are capable of performing object detection based on the provided prompts in a zero-shot manner, as illustrated in Figure 1.

However, for specific target applications such as autonomous vehicle perception [2], these foundational models may still be suboptimal. This is primarily due to the
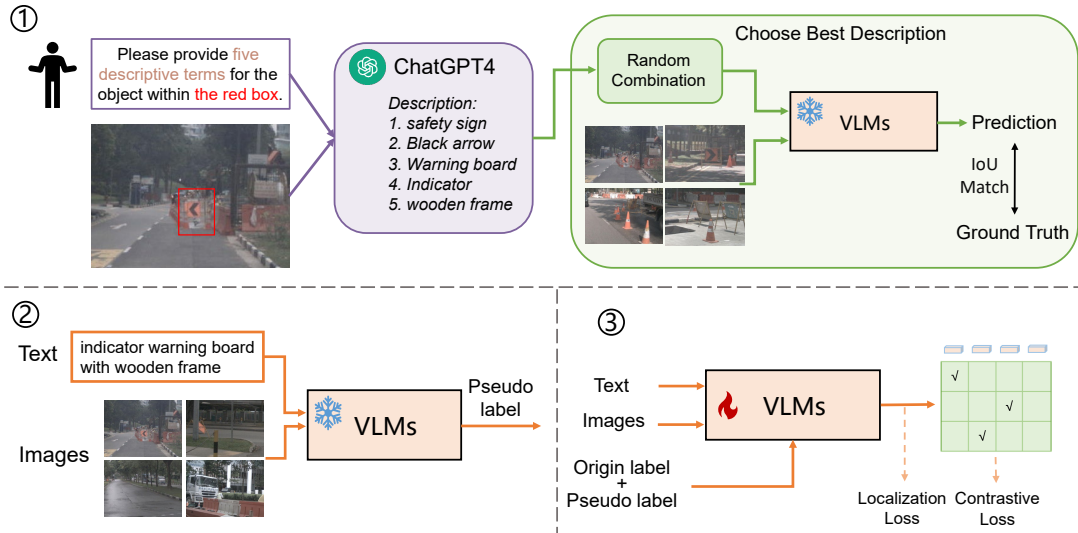
Figure 3: The framework of VLM+.

challenge of aligning the foundational models with specific target concepts as shown in Figure 2. Under the Few-Shot Object Detection setting in [5], this misalignment can have a significant impact on methods that rely on predicting pseudo-labels for images. Therefore, our goal is to reduce the gap in understanding between visual and textual concepts by the VLM, thereby minimizing the potential for generating erroneous pseudo-labels. We propose a multi-stage approach named VLM+. Specifically, we first input images with annotations for each category into a MM-LLM [10]. Here, we use GPT-4 [1] to generate keyword prompts for these categories. Then, we randomly combine these prompts as referential expressions for the vision-language object detection model to obtain an optimal referential expression for each category, thereby enhancing the foundational model's understanding of the target concepts. Finally, we utilize the acquired referential expressions as textual input to improve the generation of pseudo-labels for each category. These pseudo-labels are then employed alongside the original labeled data as annotation data for the VLMs. Additionally, the trained model can be reused for pseudo-label generation and further optimization. The object detection VLMs we utilize comprise Grounding DINO and GLIP, with the corresponding pre-trained weights available at: https://github.com/open-mmlab/mmdetection/tree/main.

2. Method

2.1. VLMs

2.1.1   GLIP

Open-set object detection is trained using existing bounding box annotations and aims to detect arbitrary classes through

language generalization. GLIP [3] considers the object detection task as a context-free phrase localization task, while phrase localization can be viewed as a context-aware object detection task. As a result, both can be improved within the same framework.

2.1.2   Grounding DINO

Grounding DINO [4] is an open-set object detector that merges the Transformer-based detector DINO with grounded pre-training. This fusion enables the detection of arbitrary objects specified by human input, like category names or referring expressions. Grounding DINO lies in its feature fusion strategy across various stages of the detection pipeline. These include feature enhancers, text-guided query selection, and cross-modal decoders, effectively integrating textual and visual information.

2.2. VLMs+

2.2.1   Concept Alignment

Regarding the misalignment between VLM and target concepts, we attribute it to the ambiguity and insufficient expression of category concepts. Therefore, to address this issue, we propose utilizing the image-to-text generation capabilities of a multimodal large language model to generate referential expressions that align with these concepts, instead of relying solely on class names, as depicted in Figure 3, labeled 1. Specifically, we overlay the annotation bounding box from the training set onto the corresponding images and set the prompt as: "Please provide five descriptive terms for the object within the red box." as input of MM-LLM. Then, we can obtain a set of descriptive prompts for each category. Leveraging the language comprehension and vision-language alignment abilities of the vision-language

Class index (c)	Class Name	Referential Expression (T_{i*}^c)	ACC (before)	ACC (after)
1	car	car	1.0	1.0
2	truck	lorry	0.6	**0.7**
3	construction vehicle	lift shovel excavator	0.9	0.9
4	bus	bus	0.9	0.9
5	trailer	large cargo box on the trailer	0.6	**0.8**
6	emergency	emergency police wagon	0.4	**0.6**
7	motorcycle	narrow motorcycle	0.8	0.8
8	bicycle	bicycle bike	0.9	**1.0**
9	adult	adult people	0.4	**0.7**
10	child	single little short youth children	0.6	**0.7**
11	police officer	traffic policeman	0.4	**0.6**
12	construction worker	construction workman people	0.5	**0.7**
13	personal mobility	small kick scooter	0.3	**0.9**
14	stroller	stroller	1.0	1.0
15	pushable pullable	pushable pullable garbage container	0.5	**1.0**
16	barrier	single short tarp barrier	0.3	**0.5**
17	traffic cone	traffic cone	1.0	1.0
18	debris	indicator warning board with wooden frame	0.0	**0.7**

Table 1: Each class name, along with its corresponding referential expression. The VLM used is Grounding DINO. **Bold** indicates improved performance.

model, we randomly combine five prompts into N referential expressions. These expressions serve as inputs to the VLM, with the aim of selecting the best referential expression for the class name. Below, we will discuss the process of selecting the optimal reference expression in detail.

Let T_i^c and $P_{i,j}^c$ denote the i-th referential expression of the c-th class and the bounding box positions of the j-th image obtained after processing T_i^c through the VLMs, respectively. B_j^c represents the the j-th ground-truth bounding box of the c-th class, where $c = 1, 2, \ldots, 18$, $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, 10$. To start, we compute the Intersection over Union (IoU) for each $P_{i,j}^c$ with B_j^c:

$$\text{IoU}(P_{i,j}^c, B_j^c) = \frac{|P_{i,j}^c \cap B_j^c|}{|P_{i,j}^c \cup B_j^c|}$$

To calculate the prediction accuracy of VLMs under the current referential expression, we first define an indicator function to check if the IoU is greater than 50%:

$$\mathbb{1}_{\text{IoU}(P_{i,j}^c, B_j^c) > 0.5} = \begin{cases} 1, & \text{if IoU}(P_{i,j}^c, B_j^c) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Next, we define the accuracy for each set of bounding boxes as:

$$\text{ACC}(P_{i,j}^c, B_j^c) = \frac{1}{10} \sum_{j=1}^{10} \mathbb{1}_{\text{IoU}(P_{i,j}^c, B_j^c) > 0.5}$$

We then find the set of bounding boxes with the highest accuracy:

$$i^* = \arg \max_{i=1,\ldots,N} \text{ACC}(P_{i,j}^c)$$

Finally, we select the referential expression T_{i*}^c as the best referential expression for c-th class name. As shown in Table 1, we present the best referential expression obtained by the VLM for each class name. Additionally, we demonstrate the improvements in prediction accuracy of VLM for each category before and after using referential expressions. We can observe significant improvements in the recognition ability for classes such as "personal mobility", "debris", "pushable pullable" and "trailer". This significantly improves the model's predictive performance on categories within a specific dataset, and holds the promise of generating higher-quality pseudo-labels for subsequent model training, thereby achieving better fine-tuning results.

2.2.2 Iterative Pseudo-label Optimization

For the federated dataset provided by the competition, we implement the iterative pseudo-label optimization approach. Iterative pseudo-label optimization involve a process where pseudo-labels, predicted labels assigned to unlabeled data by the VLMs, are iteratively generated and refined. If the confidence score of the label generated by the model for a category exceeds pseudo-label threshold η, we consider it as pseudo-labeled data for this category. Below,

we outline the detailed process of iterative pseudo-label optimization.

1. **Initial Pseudo-Label Generation**: Initially, pseudo-labels are generated for unlabeled data using the initial model and referential expressions. These pseudo-labels serve as initial labels for the unlabeled data.

2. **Model Training**: The model is then trained on both labeled data with ground truth labels and pseudo-labels. This training process aims to improve the model's performance using the combined labels.

3. **Pseudo-Label Refinement**: After training, the model's predictions on unlabeled data are updated based on the new model parameters. These updated predictions serve as refined pseudo-labels for the next iteration.

4. **Iteration**: Steps 2 and 3 are repeated iteratively, with the model being retrained on the combined labels and the pseudo-labels being refined in each iteration. This iterative process continues until a convergence criterion is met or a predefined number of iterations is reached.

2.2.3 Loss Function

The loss functions for Grounding DINO and GLIP include Focal Loss, box L1 loss, and GIOU loss. The weights for these losses are set to 1.0 for Focal Loss, 5.0 for box L1 loss, and 2.0 for GIOU loss. For Grounding DINO, similar to the DETR model, we add auxiliary losses after each decoder layer and encoder output.

Methods	mAP
Baseline (Best)	21.51
GLIP (zero-shot)	15.73
GLIP +	27.27
Grounding DINO (zero-shot)	19.91
Grounding DINO +	32.56

Table 2: Comparison of VLM and VLM+.

3. Experiment

Dataset. The train and test datasets are provided by the organizers of the Foundational Few-Shot Object Detection Challenge.

Implementation Detail. We use ChatGPT-4 to generate 5 relevant prompt descriptions for each class. For the

(a) **left**: pushable pullable; **right**: pushable pullable garbage container (33139.jpg)

(b) **left**: pushable pullable; **right**: pushable pullable garbage container (59104.jpg)

(c) **left**: personal mobility; **right**: small kick scooter (2283.jpg)

(d) **left**: personal mobility; **right**: small kick scooter (5153.jpg)

(e) **left**: debris; **right**: indicator warning board with wooden frame (15429.jpg)

(f) **left**: debris; **right**: indicator warning board with wooden frame (15937.jpg)

Figure 4: Visualizing examples: Given referential expressions about categories, VLMs can better detect new entities.

pseudo-label threshold η, we set it to 0.3. The GLIP pre-training weights are selected from: mmdetection: GLIP-L, which is pre-trained on the FourODs, GoldG, CC3M+12M,

and SBU datasets. The Grounding DINO pre-trained weights are selected from: mmdetection: MM-Grounding-DINO-L*, which is pre-trained on the O365V2, OpenImageV6, GoldG, V3det, COCO2017, LVISV1, COCO2014, GRIT, RefCOCO, RefCOCO+, RefCOCOg, and gRefCOCO datasets.

Result. To validate the effectiveness of our approach, we apply VLM+ to both the pre-trained GLIP and Grounding DINO models. As depicted in Table 2, the Grounding DINO model exhibits satisfactory performance in a zero-shot manner, owing to its extensive pre-trained on large datasets. However, as illustrated in Table 1, VLM struggles to effectively understand certain class names when they are input as text. For example, the model's prediction accuracy for the term "debris" is 0. Conversely, utilizing modified referential expressions as input significantly enhances the prediction performance for this category. The incorporation of VLM+ leads to a notable improvement in performance, showcasing the effectiveness of our approach.

Case Study. We achieve improved performance by substituting category names with referential expressions as input text for VLMs. As illustrated in Figure 4, the original terms "personal mobility" and "pushable pullable" fail to accurately capture the semantic meaning of the objects, resulting in incorrect predictions by the VLM. Moreover, for the "debris" category, the model fails to generate any predictions, indicating a very low confidence level in this category. However, utilizing enhanced referential expressions for these categories as text prompts effectively mitigates the concept misalignment issues.

4. Conclusion

This report summarize our solution for the Foundational Few-Shot Object Detection Challenge (2024). By combining MM-LLM and VLMs, and utilizing a maximum IoU matching algorithm, we identify a referential expression aligned with the image concept for each category. Subsequently, we employ iterative pseudo-label generation and model optimization under these referential expressions. The final competition results demonstrate the effectiveness of our solution.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11618–11628, 2020.

[3] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J. Hwang, K. Chang, and J. Gao. Grounded language-image pre-training. In *CVPR*, pages 10955–10965, New Orleans, LA, USA, 2022.

[4] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499, 2023.

[5] A. Madan, N. Peri, S. Kong, and D. Ramanan. Revisiting few-shot object detection with vision-language models. *CoRR*, abs/2312.14494, 2023.

[6] Y. Yang, Z. Fu, D. Zhan, Z. Liu, and Y. Jiang. Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport. *IEEE Trans. Knowl. Data Eng.*, 33(2):696–709, 2021.

[7] Y. Yang, Y. Huang, W. Guo, B. Xu, and D. Xia. Towards global video scene segmentation with context-aware transformer. In *AAAI*, pages 3206–3213. AAAI Press, 2023.

[8] Y. Yang, K. Wang, D. Zhan, H. Xiong, and Y. Jiang. Comprehensive semi-supervised multi-modal learning. In S. Kraus, editor, *IJCAI*, pages 4092–4098, 2019.

[9] Y. Yang, Y. Wu, D. Zhan, Z. Liu, and Y. Jiang. Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport. In *SIGKDD*, pages 2594–2603. ACM, 2018.

[10] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.