Simple Technical Report for the Foundational Few-Shot Object Detection Challenge 2026

Qile Miao[1], Wencan Pei[1], Zerui Xi[1], Ruijie Ma[1], Jianxin Lin[1], Yiping Gao[1]
[1] Huazhong University of Science and Technology
D202580393@hust.edu.cn, pwc@hust.edu.cn, xizerui@hust.edu.cn
M202570640@HUST.edu.cn, M202570826@hust.edu.cn, gaoyiping@hust.edu.cn

***Abstract*— *VLM-based object detectors have shown strong zero-shot capabilities on general benchmarks, yet they still struggle in realistic few-shot scenarios involving domain-specific objects, complex semantics, sparse annotations, and inconsistent labeling conventions. This report addresses the rf20vl benchmark, a challenging subset of Roboflow-VL across 20 specialized domains. We propose CoT2D, a Chain-of-Thought Driven Dual-Detector framework for few-shot object detection. CoT2D first summarizes the few annotated training examples into explicit scene, object, and annotation priors through structured prompting with Kimi-K2.6. During inference, these priors are combined with the dataset README and the test image to guide scene understanding and adaptive detector selection. The framework further integrates direct VLM-based detection with LocateAnything-based localization, balancing semantic correctness and bounding-box precision across diverse target types. Finally, a visualization-based reflection module removes obviously abnormal boxes before final prediction. Experiments and qualitative examples on rf20vl demonstrate that CoT2D provides a robust and training-free solution for detection under sparse, noisy, and semantically complex few-shot conditions.*

I. Problem Formulation

Recent VLM-based object detectors have demonstrated impressive zero-shot detection capabilities on general benchmarks such as COCO [1], largely benefiting from large-scale pretraining. These models can often detect common objects in general scenes without task-specific training. However, when facing domain-specific objects, unseen concepts, complex semantics, or professional terminology, existing vision-language models still struggle to achieve reliable zero-shot or few-shot object detection (FSOD). This limitation makes them difficult to deploy in challenging real-world scenarios where fine-tuning data and large-scale annotations are unavailable. Therefore, a central problem is how to effectively guide VLMs to inherent scene understanding and reasoning capabilities for robust object detection under strict no-fine-tuning and low-annotation constraints [2].

The Roboflow-VL subset used in this competition, referred to as rf20vl, is well aligned with this challenge [3]. It contains 20 specialized scenarios spanning medical imaging, traffic, sports, retail, biology, and other professional domains, and introduces an in-context prompt engineering track where model fine-tuning is prohibited. Beyond its domain diversity, rf20vl presents several practical challenges. First, it is a few-shot setting, where each category contains only ten annotated examples. Second, the annotations are sparse. The training and validation sets are not exhaustively labeled, meaning that they only provide positive annotations, while unlabeled regions cannot be reliably treated as negative samples. Third, the annotation style is unstable across training and test images, with inconsistent bounding-box conventions and a small number of noisy annotations. Fourth, many categories require strong global semantic understanding, where the correct label assignment depends on complex contextual reasoning rather than local appearance alone. These properties make rf20vl a realistic proxy for challenging real-world detection scenarios, where data are often collected and annotated under limited supervision and inconsistent annotation quality. Achieving strong detection performance on this benchmark therefore requires not only effective few-shot object detection capability, but also robust semantic reasoning, tolerance to noisy annotations, and strong potential for practical deployment.

II. Framework

To address the few-shot detection challenges of rf20vl while complying with the no-fine-tuning requirement of the in-context prompt engineering track, we propose a Chain-of-Thought Driven Dual-Detector (CoT2D) Framework for Few-Shot Object Detection. As shown in Fig.1, CoT2D is built upon Kimi-K2.6 [4] and formulates the detection process as a structured reasoning pipeline inspired by chain-of-thought prompting and step-wise reasoning [5], [6], [7], including training-set understanding, rule summarization, scene-level reasoning, dual-detector selection, and visualization-based reflection.

Specifically, CoT2D first constructs a chain-of-thought prompting process to extract useful information from the few annotated training examples. Instead of updating model parameters, the framework transfers the implicit scene characteristics, object appearances, category semantics, and annotation conventions contained in the training set to the detection stage through textual summaries. In this way, the limited training data are compressed into explicit contextual priors, enabling the model to better handle sparse annotations and complex semantic categories under the no-training constraint. At the detection stage, CoT2D constructs a parallel detection scheme based on both the VLM and the LocateAnything detector. According to the generality and semantic complexity of each scene, the framework adaptively selects the suitable detection detector. This design further improves annotation robustness across diverse and challenging scenarios. Finally, CoT2D introduces a visualization-based

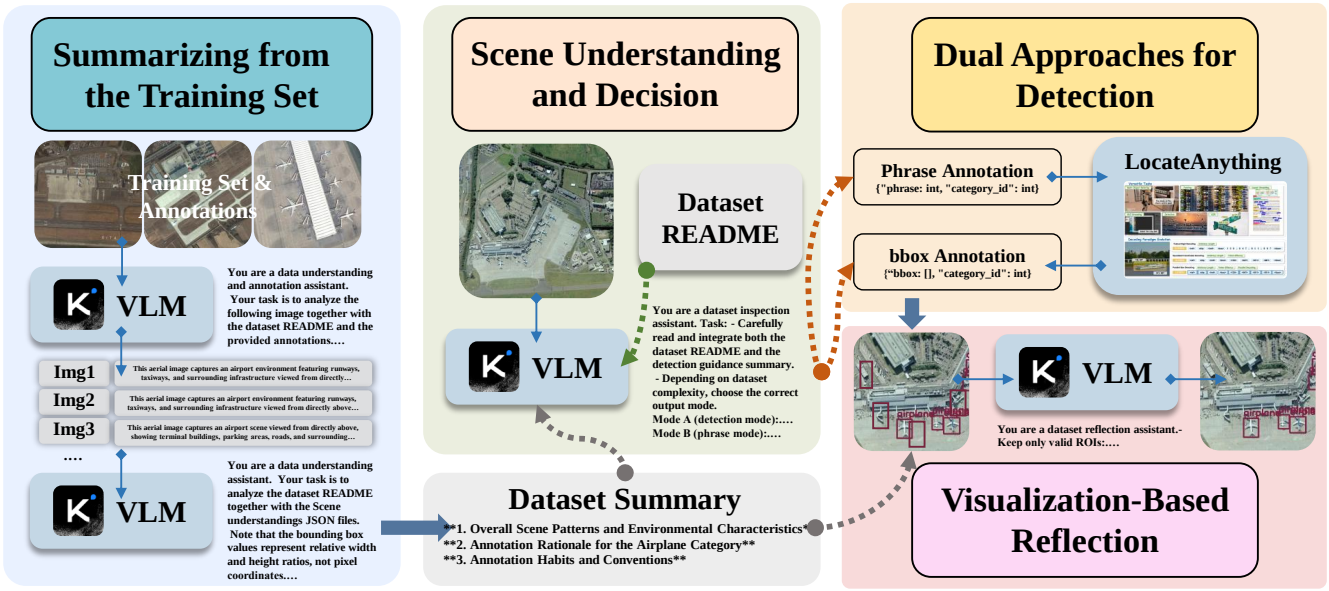

Fig. 1. Workflow of the proposed CoT2D framwork

reflection module before producing the final predictions. The initially detected bounding boxes are drawn on the input image and re-submitted to Kimi-K2.6 for a lightweight sanity check. Rather than re-detecting objects, the model inspects the visualized results and removes obviously abnormal boxes, such as boxes covering background regions, irrelevant objects, or severely misaligned areas. This reflection step further suppresses false positives and improves the stability of the final detection results.

III. Methodology of CoT2D

A. Summarizing from the Training Set

The rf20vl benchmark contains many sub-datasets that require strong global semantic understanding. For example, the volleyball dataset requires the model to infer whether the current scene corresponds to an attacking, defending, or serving phase, while the paper and webpage datasets require the model to understand the overall layout structure of the entire image. However, the information provided in the README is usually brief and cannot fully describe such dataset-specific semantic and annotation rules. Therefore, it is crucial to distill more informative scene-level, category-level, and annotation-level guidance from the few sparsely annotated training examples. Therefore, CoT2D builds a training-set summarization pipeline based on Kimi-K2.6 and prompt engineering. The pipeline follows a chain-of-thought process consisting of image understanding, key-rule summarization, and scene reasoning. In the image understanding stage, the model is prompted to describe the scene context, target object characteristics, and annotation patterns in each annotated image. This allows the framework to extract not only local object appearances, but also the surrounding context and dataset-specific labeling preferences. In the key-rule summarization stage, Kimi-K2.6 aggregates the descriptions from multiple training samples within the same sub-dataset and summarizes the common scene properties, object semantics, and annotation conventions. Since rf20vl contains sparse annotations, unlabeled regions cannot be treated as reliable negative samples. Therefore, the prompt explicitly reminds the model that the absence of annotations does not necessarily imply background, preventing it from deriving misleading negative rules from incomplete labels. The resulting summary is then used as a supplement to the original README and provided to Kimi-K2.6 during the subsequent reasoning and detection stages. Through this process, CoT2D converts a small set of sparse annotations into explicit contextual priors without any parameter update. These priors provide more reliable guidance for complex semantic understanding and category judgment in rf20vl, as illustrated by the example in Fig. 2.

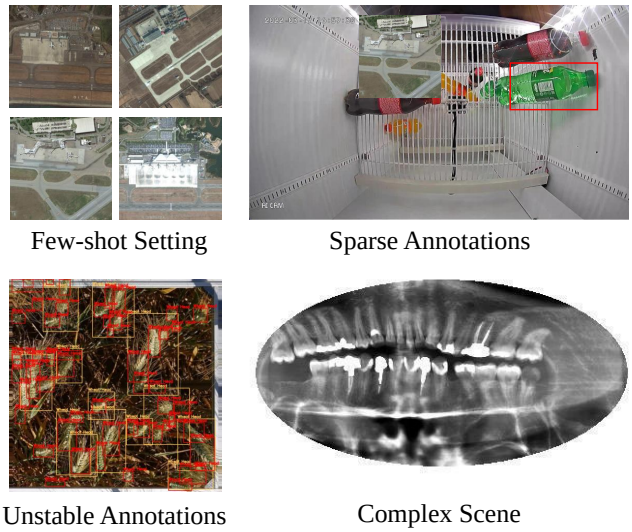

Fig. 2. Typical scenarios

B. Scene Understanding & Decision with VLM

After summarizing the training set, CoT2D further uses Kimi-K2.6 as the scene understanding and detection decision module. First, the framework sequentially feeds the training images and their annotations from the same sub-dataset into Kimi-K2.6, prompting the model to describe the scene content, target object characteristics, and bounding-box distribution patterns. Since rf20vl is both few-shot and sparsely annotated, a single image cannot provide a complete view of all possible target distributions. Therefore, CoT2D does not treat unlabeled regions as negative samples, but instead encourages the model to focus on the shared properties of annotated targets, including their appearance, spatial distribution, category semantics, and annotation style. Kimi-K2.6 then summarizes the description sequence within each sub-dataset to obtain dataset-level scene rules, object rules, and annotation rules. By explicitly summarizing these rules, CoT2D transfers useful information from the training set to the prediction stage in the form of prompts, without requiring parameter fine-tuning. This enables the model to make use of the limited annotated examples. During inference, the test image, the dataset README, and the summarized rules are jointly provided to Kimi-K2.6. Following recent multimodal chain-of-thought reasoning methods [8], the model first interprets the scene type, target semantics, and potential object regions by combining the global image content with the summarized contextual priors. It then determines the appropriate detection approaches form according to the generality and semantic complexity of the target category. Through this process of training-set rule summarization, test-image understanding, and detection-approach decision, CoT2D provides a stable interface between scene-level reasoning and the subsequent dual-detector stage under few-shot, sparse-annotation, and no-fine-tuning constraints.

C. Dual Approaches for Detection

At the detection stage, CoT2D adopts a dual-detector strategy that combines direct VLM-based detection with LocateAnything-based localization. This design is intended to handle the diverse target types in rf20vl, where different sub-datasets may require different trade-offs between semantic correctness and localization accuracy. The first detection approach is direct detection with Kimi-K2.6. Its main advantage lies in semantic understanding. By jointly considering the README, the summarized training-set rules, and the test image, Kimi-K2.6 can reason about the target category, the number of instances, and whether a candidate object should be annotated. This is particularly useful for specialized, abstract, and context-dependent scenarios, such as categories that require recognizing game phases, understanding document layouts, interpreting medical semantics, or following dataset-specific business rules. In these cases, direct VLM detection can better capture the category definition and reduce false negatives or false positives caused by ambiguous category semantics. Moreover, the VLM is strong at deciding which objects should be detected, enabling more accurate control over the predicted categories and candidate instance numbers. However, direct VLM detection also has limitations. Since its bounding boxes are not produced by a dedicated localization model, the predicted box locations may be affected by background clutter, nearby distractors, or the global image layout. As a result, it may generate shifted boxes, overly large boxes, or excessive candidate boxes. Therefore, this approach is more suitable for cases where semantic reasoning and category judgment are more important than fine-grained localization precision. The second detection approach is localization with LocateAnything [9]. When the target is a visually common object with a clear boundary, LocateAnything can generate more stable and better-aligned boxes from a concise object description, thereby improving localization accuracy. For categories with clear shapes, simple semantics, and little need for complex reasoning, this approach can compensate for the limited box precision of direct VLM detection. Nevertheless, LocateAnything also has clear limitations. As a pretrained localization model, it is more suitable for prompts composed of common object names or short visual descriptions. When the target category involves professional terminology, complex semantic rules, or context-dependent judgment, directly using the original category name as the localization prompt may lead to incorrect localization or reduced robustness. To mitigate this issue, CoT2D leverages Kimi-K2.6 to generate targeted localization phrases for LocateAnything. Specifically, the VLM interprets the category definition, the summarized training-set rules, and the test image, and then converts the target concept into a concise and visually grounded phrase that is more suitable for the localization model. This prompt-engineering step bridges the gap between complex dataset-specific category semantics and the short visual descriptions preferred by LocateAnything. By combining these two detection approaches, CoT2D establishes a complementary balance between semantic correctness and localization precision. Direct VLM detection handles complex semantics and domain-specific scenarios, while LocateAnything improves box quality for general objects. This dual-detector strategy avoids the limitations of relying on a single detector and allows the framework to better adapt to the diverse target characteristics and annotation requirements of rf20vl.

D. Visualization-Based Reflection

After the initial detection stage, CoT2D introduces a visualization-based reflection module to remove obviously unreasonable bounding boxes. Specifically, the framework draws the detected boxes on the original image and feeds the visualized result back into Kimi-K2.6 together with the category information, the README, and the summarized training-set rules. The model then inspects the displayed boxes from both the visual and semantic perspectives, identifying boxes that are clearly misaligned with the target, placed on background regions, abnormally scaled, or inconsistent with the target category. The main purpose of this module is to reduce the impact of obvious false positives on the final prediction. Due to the complex scenes and diverse category semantics in rf20vl, direct VLM detection may produce

shifted boxes, overly large boxes, or redundant candidate boxes. LocateAnything may also select visually similar background regions or irrelevant objects when the target prompt is professional or semantically complex. If all candidate boxes are directly preserved, such abnormal predictions can introduce clear false positives into the final results. Instead of performing a full re-detection, the reflection module serves as a lightweight visual sanity check. Kimi-K2.6 only examines the already visualized candidate boxes and removes those that are clearly unreasonable. In this way, the framework performs a simple but effective post-processing step without additional training or hand-crafted rules, reducing the interference of abnormal boxes and improving the stability of the final predictions. This reflection mechanism complements the previous stages. The training-set summarization module provides annotation priors, the scene understanding and dual-detector modules generate candidate results, and the reflection module performs the final visual check. By reusing the visual understanding and semantic verification capabilities of the VLM, CoT2D further improves detection robustness under few-shot and sparse-annotation conditions.

IV. Performance of CoT2D

We evaluate CoT2D on the rf20vl benchmark under the no-fine-tuning in-context prompt engineering setting, as shown in Fig. 3. By combining training-set rule summarization, adaptive dual-detector selection, and visualization-based reflection, CoT2D shows robust performance across diverse domains, including medical imaging, sports, traffic, retail, biology, documents, and webpages. These results demonstrate that the proposed framework can effectively handle few-shot detection with sparse annotations, complex category semantics, and inconsistent labeling conventions.

Fig. 3. Visualization result of CoT2D

V. Conclusion

This report presents CoT2D, a training-free prompt engineering framework for few-shot object detection on the challenging rf20vl benchmark. Instead of relying on parameter fine-tuning, CoT2D converts limited and sparsely annotated examples into explicit contextual priors, enabling the VLM to better understand dataset-specific scene semantics, object definitions, and annotation conventions. By combining scene-aware detector selection, complementary VLM and LocateAnything detection paths, and visualization-based reflection, the framework improves robustness against semantic ambiguity, localization errors, and abnormal false positives. The results suggest that carefully designed in-context reasoning can serve as an effective alternative to fine-tuning when large-scale annotations are unavailable. More broadly, CoT2D highlights the potential of prompt-driven visual reasoning for practical object detection in specialized, low-resource, and imperfectly annotated real-world scenarios.

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[2] A. Madan, N. Peri, S. Kong, and D. Ramanan, "Revisiting few-shot object detection with vision-language models," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 19 547–19 560.

[3] P. Robicheaux, M. Popov, A. Madan, I. Robinson, J. Nelson, D. Ramanan, and N. Peri, "Roboflow100-vl: A multi-domain object detection benchmark for vision-language models," 2025. [Online]. Available: https://arxiv.org/abs/2505.20612

[4] Moonshot AI, "Kimi k2.6," https://huggingface.co/moonshotai/Kimi-K2.6, 2026, accessed: 2026-06-01.

[5] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 22 199–22 213.

[6] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi, "Least-to-most prompting enables complex reasoning in large language models," in *International Conference on Learning Representations*, 2023.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24 824–24 837.

[8] C. Mitra, B. Huang, T. Darrell, and R. Herzig, "Compositional chain-of-thought prompting for large multimodal models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 420–14 431.

[9] S. Wang, S. Liu, Y. Kuang, X. Wei, Y. Liu, Z. Li, Y. Man, G. Chen, A. Tao, G. Liu, J. Kautz, L. Zhang, and Z. Yu, "Locateanything: Fast and high-quality vision-language grounding with parallel box decoding," *arXiv:2605.27365*, 2026.