VLM-Guided Detection and Rematching for Private Object Localization

Heeseung Cho[1] Yuna Park[1] Esther Kim[2] Jiho Kim[2] Hyojoo Kim[2]
Christian Wallraven[1,3] Junhyoung Oh[2*]

[1]Department of Artificial Intelligence, Korea University
[2]Division of Information Security, Seoul Women's University
[3]Department of Brain and Cognitive Engineering, Korea University

Abstract

We present a training-free framework for private object localization. Rather than fine-tuning a detector on the target domain, we retain an off-the-shelf open-vocabulary detector (OVD) and align it through vision-language model (VLM)-guided semantic prompts. Given any query image, our approach operates in two phases: (i) a VLM generates category-aware cues and identifies a PII-type supercategory to guide Grounding DINO toward privacy-relevant regions, and (ii) the same VLM jointly re-evaluates all detected candidates on the full image, assigning categories within a restricted candidate space and rejecting false positives. SAM produces the final segmentation masks. Experiments on the Biv-Priv-Seg dataset demonstrate strong performance without any model fine-tuning or support augmentation, surpassing the 2025 top-performing method LLM2Seg by +3.64 bbox mAP and +3.68 segmentation mAP. Our code is available at https://github.com/Heeseung-Cho/VLM-Guided-FSOL.

1. Introduction

The VizWiz FSOL challenge [3] targets the localization of 16 private object categories in images taken by blind and low vision (BLV) users. Participants receive a support set of 16 labeled images (one per category) and a query set of 1,056 unlabeled images, and must predict bounding boxes and segmentation masks without external training data. Prior work [3] reports that open-vocabulary detector (OVD) reached only AP_{50} 60.2 even when given category-specific prompts. This ceiling suggests that providing accurate category names alone is not sufficient for OVDs to localize these private objects.

Instead of fine-tuning the detector on the target domain, we propose a framework that augments an off-the-shelf OVD with vision-language model (VLM)-guided semantic prompts and candidate rematching, requiring no fine-tuning or support image references.

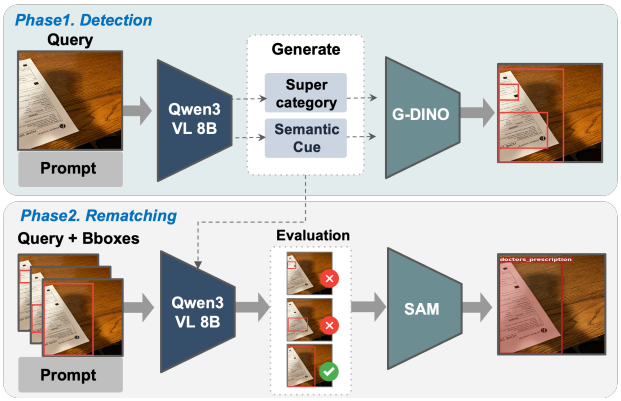

Figure 1. Overview of our proposed method.

2. Method

Figure 1 shows our framework performing concept alignment in two phases around an OVD (Grounding DINO).

Phase 1: Detection. A VLM (Qwen3-VL-8B) examines the query image and produces two outputs. First, the VLM generates ***semantic cues***: detector-friendly noun phrases that describe the target beyond its category name. Second, it predicts a ***supercategory (SC)*** among four predefined groups based on the PII taxonomy [2]: *document* (8 categories), *health* (5), *card* (2), and *tattoo* (1). All categories within the predicted group are included as detection targets alongside the semantic cues, and passed as text prompts to OVD, which generates candidate bounding boxes over a wider search space than fixed category names alone.

Phase 2: Rematching. The query image, annotated with all proposed bounding boxes, is presented to the VLM in a single call together with the Phase-1 context and a scene caption. The VLM rematches each open-vocabulary proposal against the closed set of categories within the predicted SC—accepting boxes that correspond to a target cate-

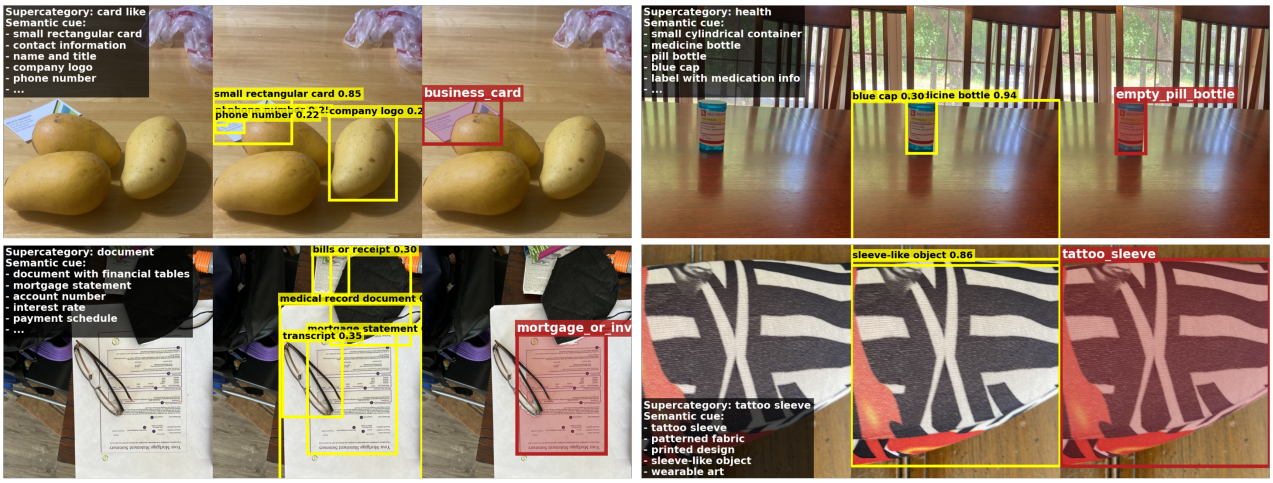

Figure 2. Pipeline examples for each supercategory. (Left) Phase-1 VLM output. (Middle) Candidates produced from Phase-1 prompt. (Right) Final predictions after Phase-2, which rejects false positives and assigns the correct category.

	Ablation	bbox mAP	seg mAP
tele-ai-a [4]		59.00	–
LLM2Seg [5]		61.63	60.14
Ours	–	56.34	54.69
	+ Cue	63.44	61.85
	+ SC	64.87	63.19
	+ Cue + SC	**65.27**	**63.82**

Table 1. Ablation and comparison on the challenge.

gory and rejecting the rest. Accepted predictions are passed to SAM for segmentation.

3. Results

Tab. 1 shows that our method achieves 65.27 bbox mAP and 63.82 segmentation mAP without any fine-tuning, outperforming the prior winner LLM2Seg [5] by +3.64 and +3.68 respectively. We ablate two components: semantic cue (Cue) and SC effect. The baseline—where the VLM passes only predicted category names to the detector—achieves 56.34 mAP. Cue adds +7.10 and SC adds +8.53 independently; combining both yields +8.93 over the baseline. Fig. 2 illustrates the full pipeline on examples from different SC.

4. Discussion and Conclusion

Our results confirm that the detector ceiling reported in [3] can be addressed through semantic alignment rather than detector adaptation. Jointly rematching open-vocabulary proposals against a closed category set [1] is essential—it is the only stage where false positives are rejected and categories are reassigned with full image context. We retain the pretrained detector without fine-tuning, and the VLM operates without support image references, as neither adaptation is necessary in our framework.

Our work shows that aligning an off-the-shelf detector with VLM-generated semantic prompts is a viable alternative to detector fine-tuning for BLV private object localization.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25476234).

References

[1] 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

[2] Jeffri Murrugarra-Llerena, Niu Haoran, Barber K.Suzanne, Hal Daume III, Yang Trista Cao, and Paola Cascante-Bonilla. Beyond blanket masking: Examining granularity for privac protection in images captured by blind and low vision users. *COLM*, 2025. 1

[3] Yu-Yun Tseng, Tanusree Sharma, Lotus Zhang, Abigale Stangl, Leah Findlater, Yang Wang, and Danna Gurari. Bivpriv-seg: Locating private content in images taken by people with visual impairments. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 430–440, 2025. 1, 2

[4] VizWiz. Vizwiz private objects localization 2024. *VizWiz Grand Challenge Workshop at CVPR*, 2024. 2

[5] Wei-Chih Yin, Pin-Hsuan Chou, Chao-Chi Liao, Yu-Chee Tseng, and Cheng-Kuan Lin. Llm2seg: Llm-guided few-shot object localization with visual transformer. *VizWiz Grand Challenge Workshop at CVPR*, 2025. 2