TOWARDS LESS BASE-BIASED OPEN-VOCABULARY OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen-backbone detectors have emerged as an effective paradigm for Open-Vocabulary Object Detection (OVOD), as they preserve transferable knowledge from pretrained vision-language models. However, freezing the backbone alone does not eliminate detector-level base bias: base-only optimization can distort novel-region representations, while region proposal networks trained on base categories often produce spatially inaccurate proposals for unseen objects. To address these issues, we propose **ToLB**, a framework **To**wards **L**ess base-**B**iased OVOD that tackles representation and localization bias at training and inference, respectively. Specifically, we propose a Relational Knowledge Preservation module that employs a frozen visual foundation model as a category-agnostic structural teacher and regularizes region representations through pairwise relational consistency, avoiding restrictive alignment between heterogeneous feature spaces. At inference, we further introduce a VLM-Guided Region Refinement module that extracts category-agnostic objectness cues from dense vision-language features to identify foreground-dominant regions and adaptively refine proposals, thereby reducing background interference for novel objects. Extensive experiments on OV-COCO and OV-LVIS demonstrate superior performance in novel-category generalization, achieving 48.2% Novel AP_{50} on OV-COCO and 39.5% Rare mask mAP on OV-LVIS, respectively.

1 INTRODUCTION

Open-Vocabulary Object Detection (OVOD) (Zareian et al., 2021) aims to detect both base categories seen during training and novel categories specified only at inference. Existing methods mainly improve open-vocabulary generalization through enhanced alignment strategies, including region-word alignment (Lin et al., 2023), region-of-bag alignment (Wu et al., 2023), and clustering-based region-word alignment (Liang & Zhang, 2025), or by generating pseudo-labels for novel categories during the training phase (Zhao et al., 2022; Wang et al., 2024; Pham et al., 2024). However, the former often relies on large-scale image-caption data (*e.g.*, COCO Caption (Chen et al., 2015) or CC3M (Sharma et al., 2018)) for region-language supervision, while the latter typically assumes access to predefined novel-category vocabularies for pseudo-label generation (Zhang et al., 2025).

To reduce such reliance on additional supervision and category priors, F-VLM (Kuo et al., 2023) pioneered a frozen-backbone paradigm for OVOD, where the pretrained vision-language backbone (Radford et al., 2021) is kept fixed and only detector-specific components (*e.g.*, RPN, and RoI Head (Ren et al., 2015)) are optimized on base categories with annotations. This frozen-backbone design preserves the transferable knowledge encoded in the pretrained model while adapting detection-specific components to region-level recognition. However, F-VLM primarily builds upon ResNet-based CLIP backbones, whose convolutional features are less flexible for dense region representation than their ViT counterparts (Wu et al., 2024; Wang et al., 2025). To better leverage ViT-based CLIP for dense region recognition, CLIPSelf (Wu et al., 2024) transfers image-level semantic knowledge to local regions through self-distillation, yielding stronger region-aware representations, as shown in Figure 1(a). The adapted backbone is then frozen for open-vocabulary detector training on base categories, as shown in Figure 1(b).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

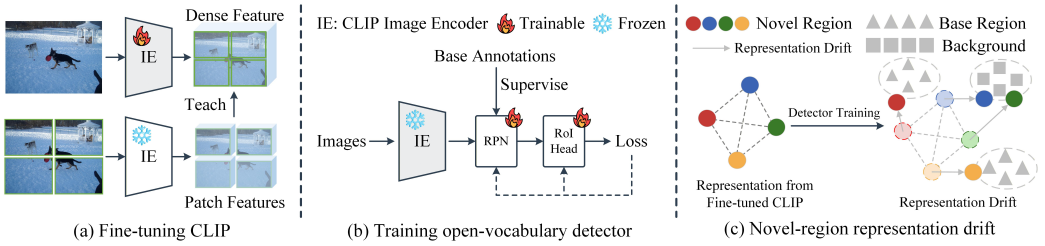

Figure 1: (a) CLIPSelf transfers image-level CLIP knowledge to dense region representations. (b) The fine-tuned backbone is frozen during open-vocabulary detector training, while the RPN and RoI head are optimized. (c) Base-category supervision can nevertheless distort the relational structure of novel-region representations and bias them toward base categories or backgrounds.

Although CLIPSelf (Wu et al., 2024) provides a stronger region-aware backbone, freezing the adapted encoder cannot eliminate bias in the downstream detector. The trainable detection components are still optimized exclusively with base-category supervision, which biases region representations toward base categories or backgrounds. As illustrated in Figure 1(c), this optimization progressively distorts the relational structure encoded by the frozen backbone, leading to novel-region representation drift. We quantify this effect in Table 1 by measuring the relational consistency (Pearson correlation) between frozen CLIPSelf and detector-level region representations. Our method increases the Pearson correlation from 0.289 to 0.354 over all categories and from 0.357 to 0.459 on novel categories. These results indicate that freezing the backbone alone is insufficient to preserve transferable relational structure during detector adaptation. Moreover, such representation bias can be further amplified by imperfect proposal localization. Since the RPN is trained only on base categories, its learned objectness and localization priors are biased toward seen objects or backgrounds, often producing spatially inaccurate proposals for novel categories (Ma et al., 2023; Wu et al., 2023; Liang & Zhang, 2025). Such proposals contain excessive background context and consequently contaminate the RoI representations used for open-vocabulary object detection.

Table 1: Relational structure preservation measured by the Pearson correlation between fine-tuned CLIP and base-trained detector representations.

Method	**Overall**	**Novel**
CLIPSelf (Wu et al., 2024)	0.289	0.357
CLIPSelf *w.* our RKP	**0.354**	**0.459**
Δ	+0.065	+0.102

The above observations motivate us to preserve the transferable region structure during training while recovering more object-centric regions at inference. In this paper, we propose **ToLB**, a framework for less base-biased open-vocabulary object detection. Specifically, during training, a Relational Knowledge Preservation (RKP) module is proposed that employs a frozen visual foundation model (*i.e.*, DINOv2 (Oquab et al., 2024)) as a category-agnostic structural teacher to preserve transferable inter-region relations. Instead of directly aligning heterogeneous feature spaces, RKP regularizes detector representations through relational consistency, mitigating representation drift induced by base-category supervision. Meanwhile, at inference, a VLM-Guided Region Refinement (VRR) module derives category-agnostic objectness cues from dense vision-language features to identify foreground-dominant regions and adaptively refine proposals, thereby suppressing background interference. Extensive experiments on standard OVOD benchmarks demonstrate consistent improvements of our ToLB in novel-category generalization.

To sum up, our contributions are summarized as follows: 1) We identify novel-region representation drift in open-vocabulary detectors as an overlooked limitation of the frozen-backbone paradigm and propose ToLB to alleviate this bias. 2) A relational knowledge preservation module is introduced to preserve transferable inter-region relations and mitigate representation drift induced by base-category supervision. 3) A VLM-guided region refinement module is developed to adaptively refine novel-object proposals using category-agnostic objectness cues, thereby reducing background interference. 4) Extensive experiments on OV-COCO and OV-LVIS demonstrate consistent improvements in novel-category generalization, with ToLB achieving 48.2% Novel AP_{50} on OV-COCO and 39.5% Rare mask mAP on OV-LVIS.

2 Related Work

Open-vocabulary object detection extends conventional object detection beyond the categories annotated during training by transferring semantic knowledge from vision-language models to novel categories. Existing OVOD methods mainly improve novel-category generalization through three complementary directions: caption- or VLM-guided knowledge transfer, pseudo-labeling, and frozen-backbone adaptation.

2.1 Caption- or VLM-guided knowledge transfer

Image-caption supervision and pretrained vision-language models provide rich semantic knowledge beyond box-annotated base categories, enabling detectors to associate localized visual regions with a broader vocabulary. VLDet (Lin et al., 2023) formulates object-language alignment as a matching problem between region features and words, allowing the detector to learn directly from image-text pairs without additional grounding annotations. BARON (Wu et al., 2023) goes beyond independent region matching by grouping contextually related regions into bags and aligning their compositional representations with frozen VLM features. CADet (Liang & Zhang, 2025) further incorporates SAM (Kirillov et al., 2023) to enrich novel-object proposals and clusters semantically similar regions to recover additional region-word correspondences.

However, these approaches rely on additional semantic supervision or auxiliary pretrained models to establish reliable region-language correspondence, making their performance dependent on the quality and coverage of such external knowledge.

2.2 Pseudo-labeling based knowledge transfer.

Pseudo-labeling provides an effective way to expose detectors to unlabeled objects beyond the base categories, thereby enriching region-level supervision without requiring exhaustive bounding-box annotations. VL-PLM (Zhao et al., 2022) generates pseudo labels on unlabeled images by combining class-agnostic region proposals with CLIP to localize and categorize candidate objects. MarvelOVD (Wang et al., 2024) further improves VLM-generated pseudo labels by jointly exploiting detector and VLM predictions. LP-OVOD (Pham et al., 2024) retrieves pseudo labels from region proposals relevant to novel text queries and trains a lightweight linear classifier to distinguish high-quality object regions from poorly localized proposals.

Despite their effectiveness, many pseudo-labeling methods rely on prior access to novel-category vocabularies, limiting their applicability to realistic open-world settings. Moreover, the generated pseudo labels can be noisy or spatially inaccurate due to the gap between image-level vision-language pretraining and object-level localization.

2.3 Frozen Feature Backbones for OVOD

Freezing pretrained vision-language backbones provides an effective means of preserving open-vocabulary knowledge during detector training, avoiding the degradation of transferable semantics caused by task-specific fine-tuning. CLIPSelf (Wu et al., 2024) adapts image-level CLIP knowledge to local regions through self-distillation, aligning dense region features with the representations of their corresponding image crops without requiring region-text annotations. Building upon this paradigm, DeCLIP (Wang et al., 2025) decouples CLIP representations into content and context components, enhancing local discriminability while preserving spatial correlations under the guidance of visual foundation models. More recently, LLCS (Liang & Zhang, 2026a) further improves the locality and semantic consistency of dense VLM representations, yielding stronger transferable features for downstream open-vocabulary object detection.

Despite increasingly strong region-aware backbones, existing methods primarily preserve transferable knowledge at the backbone level. However, freezing the visual encoder does not constrain how downstream detection components reorganize and localize this knowledge under base-category supervision. As a result, novel-region representations can drift toward base-category or background-dominated decision regions, while base-trained proposal generators remain biased in localizing novel objects. To overcome these issues, in this paper, we extend knowledge preservation beyond the backbone, targeting both detector-level representation drift and proposal bias.

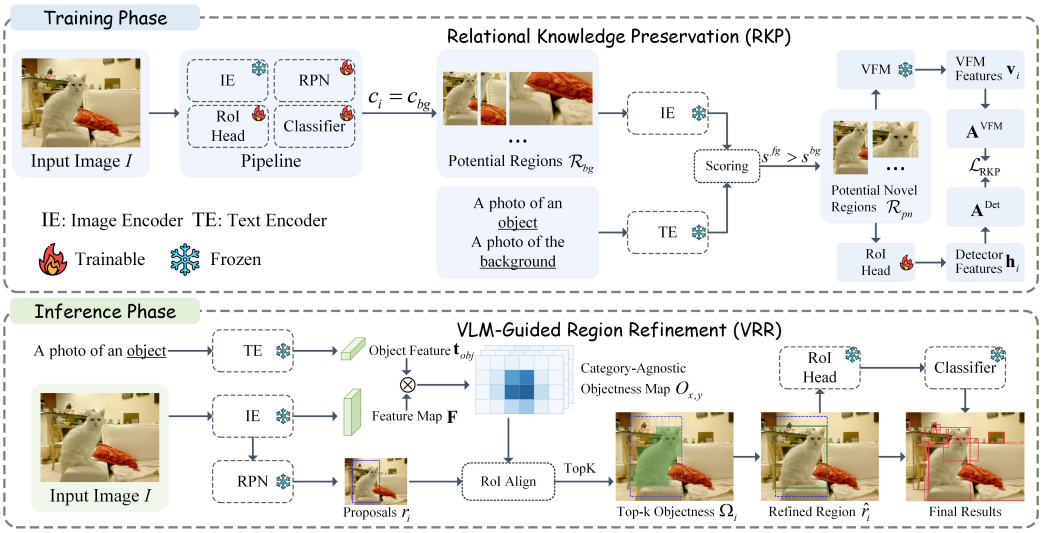

Figure 2: Overview of the proposed ToLB. Relational Knowledge Preservation (RKP) preserves transferable relational structure for potential novel regions during training, while VLM-Guided Region Refinement (VRR) exploits category-agnostic VLM objectness to refine novel-object proposals at inference. For clarity, the bounding-box regression branch is omitted.

3 METHOD

3.1 PRELIMINARIES

Let $\mathcal{C}_B$ and $\mathcal{C}_N$ denote the disjoint sets of base and novel categories, respectively. During training, only instances from $\mathcal{C}_B$ are annotated, while the detector is required to recognize $\mathcal{C}_B \cup \mathcal{C}_N$ at inference. Following F-VLM (Kuo et al., 2023) and CLIPSelf (Wu et al., 2024), we adopt a frozen-backbone two-stage detector. Given an image I, the frozen encoder $\mathcal{E}_v$ extracts dense visual features $\mathbf{F} = \mathcal{E}_v(I)$, and the trainable RPN generates region proposals $\{r_i\}_{i=1}^{N}$. The corresponding region representations are obtained by the RoI head from RoI-aligned features. For open-vocabulary recognition, category names are encoded into text prototypes (*e.g.*, *A photo of a/an [category name]*) by the frozen CLIP text encoder and matched with region representations in the shared vision-language space. During detector training, the visual backbone remains fixed, while the RPN and RoI head are optimized exclusively with base-category supervision.

As discussed in Section 1, although this paradigm preserves transferable knowledge in the pretrained backbone, it does not prevent base-category bias from emerging in the downstream detector. In other words, base-only optimization can distort novel-region representations and produce spatially inaccurate proposals for unseen objects. These two limitations motivate us to propose the ToLB framework for less base-biased OVOD, which includes Relational Knowledge Preservation (RKP) and VLM-Guided Region Refinement (VRR) to address detector-level representation drift and proposal-level spatial bias, respectively.

3.2 OVERVIEW

We propose ToLB, a framework for mitigating detector-level base bias in frozen-backbone open-vocabulary object detection. As illustrated in Figure 2, ToLB addresses two complementary failure modes in frozen-backbone OVOD. During training, Relational Knowledge Preservation (RKP) is proposed to identify potential novel regions that are mistakenly treated as backgrounds and to preserve their transferable relational structure using a frozen visual foundation model. At inference, VLM-Guided Region Refinement (VRR) exploits category-agnostic objectness cues from dense vision-language features to recover foreground-dominant regions and refine spatially biased proposals. Detailed descriptions of the proposed components can be found in the following.

3.3 RELATIONAL KNOWLEDGE PRESERVATION (RKP)

Although the image encoder remains frozen, the RPN, RoI head, and classifier are optimized solely with base-category supervision, as shown in the pipeline of Figure 2. As a result, unlabeled novel objects are implicitly treated as negatives during detector training and may therefore be increasingly suppressed as background, even when their spatial extent is reasonably captured by the proposal generator. To quantify this effect, we measure the proportion of novel-category ground-truth instances that are overlapped by RoI proposals with IoU > 0.5 but are assigned as background during base-category training. As shown in Figure 3, this proportion increases from 39.1% in the first epoch to 44.6% and 45.6% in the second and third epochs, respectively. This trend indicates that continued optimization under base-only supervision progressively strengthens the suppression of unlabeled novel regions at the detector level. Consequently, although the frozen backbone retains transferable visual knowledge, the downstream detector increasingly distorts this knowledge when forming region representations for novel objects. [1]

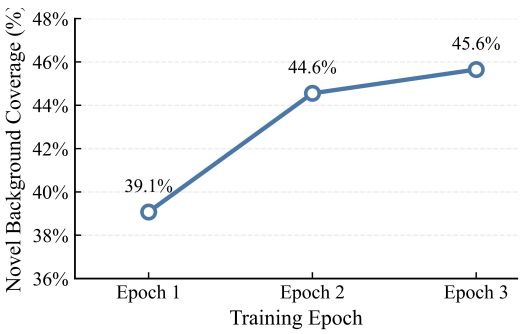

Figure 3: Novel-object background coverage during base-category training on OV-COCO. We report the percentage of novel GT instances covered by RoI proposals assigned as background with IoU > 0.5 in each training epoch.

Motivated by this observation, we introduce Relational Knowledge Preservation (RKP). RKP first reassesses detector-background regions using the category-agnostic semantics of a frozen VLM to identify potential novel regions, and then preserves their transferable relational structure using a frozen visual foundation model (Oquab et al., 2024).

Potential novel region mining. Let $\mathcal{R} = \{r_i\}_{i=1}^{N}$ denote the proposals generated by RPN. Each proposal is first pooled from the shared feature maps by RoIAlign, then transformed into a detector-level region representation by the RoI head, and finally fed into the classifier for category prediction. We collect sampled RoI regions assigned to the background under base-category supervision to form the potential regions set, as illustrated in the training phase of Figure 2, $\mathcal{R}_{bg} = \{r_i \in \mathcal{R} \mid c_i = c_{bg}\}$, where c_i denotes the training category label assigned to region r_i, and c_{bg} denotes the background class. As suggested by Figure 3, $\mathcal{R}_{bg}$ may contain a considerable number of unlabeled novel objects. We therefore leverage the open-vocabulary semantics of the frozen VLM (*i.e.*, CLIP) to identify regions exhibiting foreground characteristics. Since the VLM is adapted beforehand and remains frozen during detector training, its representations are not directly affected by the base-category optimization of the detector, making them less susceptible to detector-induced base bias.

Specifically, we encode two category-agnostic prompts, "a photo of an object" and "a photo of the background", using the frozen text encoder to obtain normalized text embeddings $\mathbf{t}_{fg}$ and $\mathbf{t}_{bg}$. For each region $r_i \in \mathcal{R}_{bg}$, we extract its regional VLM representation $\mathbf{z}_i$ from the frozen dense visual features and compute its foreground and background semantic similarities as:

$$s_i^{fg} = \frac{\mathbf{z}_i^\top \mathbf{t}_{fg}}{\|\mathbf{z}_i\|_2 \|\mathbf{t}_{fg}\|_2}, \qquad s_i^{bg} = \frac{\mathbf{z}_i^\top \mathbf{t}_{bg}}{\|\mathbf{z}_i\|_2 \|\mathbf{t}_{bg}\|_2}. \tag{1}$$

The scoring operation in Figure 2 corresponds to the process defined above. A detector-background region is regarded as a *potential novel region* when its generic foreground evidence exceeds its background evidence, which is formulated as:

$$\mathcal{R}_{pn} = \left\{ r_i \in \mathcal{R}_{bg} \mid s_i^{fg} > s_i^{bg} \right\}. \tag{2}$$

Our RKP does not assign any semantic category to these regions. The VLM is used only to determine whether a detector-background proposal is more likely to contain a generic foreground object than background. Therefore, the mining process requires neither novel-category annotations nor a predefined novel vocabulary.

[1] Novel-category annotations are used solely for this diagnostic analysis and are never accessed for model optimization or potential novel-region mining.

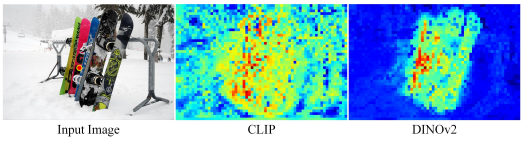

Figure 4: Token-level relational responses of CLIP and DINOv2.

Relational structure preservation. The identified potential novel regions are then used to preserve transferable structural knowledge during detector optimization. A key design choice is the selection of the structural teacher. Although the frozen CLIP representation is effective for semantic discovery, RKP aims to preserve category-agnostic inter-region structure rather than category-level semantic alignment. As illustrated in Figure 4, given the same queried region, we visualize its cosine similarity to all other patch tokens in CLIP and DINOv2. CLIP produces relatively dispersed responses, whereas DINOv2 exhibits more spatially coherent responses over structurally related object regions, suggesting a stronger structural prior for modeling inter-region relationships. This qualitative observation is further supported by our ablation study in Table 5, where replacing DINOv2 with CLIP as the relational teacher leads to inferior performance. We therefore adopt DINOv2 as the frozen VFM for relational knowledge preservation.

Specifically, given the potential novel regions $\mathcal{R}_{pn} = \{r_i\}_{i=1}^{K}$, we extract their representations from the frozen VFM and the trainable detector: $\mathbf{v}_i = \Phi_{\text{VFM}}(I, r_i), \mathbf{h}_i = \Phi_{\text{Det}}(I, r_i)$, where $\mathbf{v}_i$ denotes the VFM representation and $\mathbf{h}_i$ denotes the detector-level representation produced by the RoI head. Directly aligning $\mathbf{v}_i$ and $\mathbf{h}_i$ is undesirable because the VFM and detector are optimized with different objectives and reside in heterogeneous feature spaces. Instead, we preserve the relative geometry among potential novel regions. For each pair (r_i, r_j), we define their pairwise relations as:

$$A_{ij}^{\text{VFM}} = \frac{\mathbf{v}_i^\top \mathbf{v}_j}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2}, \qquad A_{ij}^{\text{Det}} = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\|_2 \|\mathbf{h}_j\|_2}. \tag{3}$$

The resulting matrices $\mathbf{A}^{\text{VFM}}$ and $\mathbf{A}^{\text{Det}}$ describe the relational geometry among potential novel regions in the two representation spaces. RKP encourages the relational structure of the detector to remain consistent with the category-agnostic structure provided by the frozen VFM. Accordingly, we minimize the discrepancy between the two relation matrices using the Smooth L1 loss:

$$\mathcal{L}_{\text{RKP}} = \text{SmoothL1}\left(\mathbf{A}^{\text{Det}}, \mathbf{A}^{\text{VFM}}\right). \tag{4}$$

The diagonal entries, corresponding to trivial self-relations with similarity 1, are masked out by setting them to 0 before computing the loss. By preserving inter-region relations rather than directly matching heterogeneous feature vectors, RKP constrains how potential novel regions are reorganized during base-category optimization while retaining the detector's task-specific adaptation capability. The complete procedure of RKP is summarized in Algorithm 1 in the Appendix A.1.

3.4 VLM-Guided Region Refinement (VRR)

Although RKP mitigates representation drift during training, the base-trained RPN can still produce spatially biased proposals for novel objects (Ma et al., 2023; Liang & Zhang, 2026b). These proposals often contain excessive background context and degrade region-level recognition. To address this issue, we further introduce VLM-Guided Region Refinement (VRR), which uses category-agnostic VLM objectness cues to recover accurate object-centric regions at inference.

Specifically, we encode the generic prompt "a photo of an object" with the frozen text encoder to obtain the normalized object embedding $\mathbf{t}_{obj}$. Given the dense visual feature map $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ from the frozen image encoder, we construct a category-agnostic objectness map by:

$$O_{x,y} = \frac{\mathbf{F}_{x,y}^\top \mathbf{t}_{obj}}{\|\mathbf{F}_{x,y}\|_2 \|\mathbf{t}_{obj}\|_2}. \tag{5}$$

This generic objectness cue provides spatial guidance without accessing novel-category semantics. Then, for each RPN proposal r_i, RoIAlign extracts its proposal-level objectness response:

$$\mathbf{S}_i = \text{RoIAlign}\left(\mathbf{O}, r_i\right), \qquad \mathbf{S}_i \in \mathbb{R}^{R \times R}. \tag{6}$$

We retain the k locations with the highest responses as the objectness region, formulated as:

$$\Omega_i = \text{TopK}\left(\mathbf{S}_i, k\right). \tag{7}$$

Ω_i identifies the foreground-dominant area within the proposal. Finally, we obtain the refined region from the spatial extent of these selected locations, which are formulated as:

$$\hat{r}_i = \left[\min_{j \in \Omega_i} x_j, \min_{j \in \Omega_i} y_j, \max_{j \in \Omega_i} x_j, \max_{j \in \Omega_i} y_j \right], \tag{8}$$

where $\hat{r}_i$ denotes the final object-centric proposal. Compared with the original proposal (r_i), $\hat{r}_i$ retains spatial regions that are strongly supported by the VLM while suppressing irrelevant background context. Finally, the refined proposals are pooled by RoIAlign and passed through the RoI head and open-vocabulary classifier for recognition. The complete VRR procedure is provided in Algorithm 2 of Appendix A.2.

3.5 Training Objective

During training, the pretrained image encoder and the VFM teacher remain frozen, while the detector is optimized using the standard CLIPSelf detection objective along with the proposed RKP loss:

$$\mathcal{L} = \underbrace{\mathcal{L}^{\text{cls}}_{\text{RPN}} + \mathcal{L}^{\text{reg}}_{\text{RPN}} + \mathcal{L}^{\text{cls}}_{\text{RoI}} + \mathcal{L}^{\text{reg}}_{\text{RoI}}}_{\mathcal{L}_{\text{CLIPSelf}}} + \lambda_{\text{RKP}} \mathcal{L}_{\text{RKP}}, \tag{9}$$

where $\mathcal{L}_{\text{CLIPSelf}}$ consists of the classification and box regression losses of RPN and RoI head, and λ_{RKP} controls the trade-off between detection optimization and relational knowledge preservation.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation metrics. We evaluate ToLB on two standard open-vocabulary object detection benchmarks, OV-COCO (Lin et al., 2014) and OV-LVIS (Gupta et al., 2019). For OV-COCO, following the widely adopted split (Zareian et al., 2021), 48 categories are used as base classes, and 17 categories are held out as novel classes. We report AP_{50} on novel, base, and all categories, denoted as Novel AP^N_{50}, Base AP^B_{50}, and All AP^{All}_{50}. For OV-LVIS, we follow the standard open-vocabulary setting (Gu et al., 2022) that treats the 337 rare categories as novel classes and the remaining common and frequent categories as base classes. We report mask AP on rare, common, frequent, and all categories, denoted as mAP^{mask}_r, mAP^{mask}_c, mAP^{mask}_f, and mAP^{mask}_{all}.

Implementation details. We build ToLB on a frozen-backbone two-stage open-vocabulary detector following CLIPSelf (Wu et al., 2024), using ViT-B/16 and ViT-L/14 as backbone variants. The region-aware visual backbone remains frozen throughout training, while the RPN and RoI head are optimized using only base-category annotations. For the relational knowledge preservation module, CLIP is used to mine potential novel regions from detector-background predictions, while frozen DINOv2-ViT-B/14 and DINOv2-ViT-L/14 (Oquab et al., 2024) models serve as structural teachers for the corresponding backbone variants. We set $\lambda_{\text{RKP}} = 5.0$ in Eq. 9 for all experiments, with a sensitivity analysis in Appendix A.3. In the VLM-guided region refinement module, for each proposal, dense VLM responses are pooled into an 7×7 grid, and the Top 40% in Eq. 7 foreground-responsive locations are retained to refine its spatial extent, with detailed hyper-parameter settings provided in the Appendix A.4. Unless otherwise specified, all other detector configurations and training schedules follow the baseline. All experiments are conducted on 8 NVIDIA L20 GPUs.

4.2 Quantitative Results

Results on OV-COCO. Table 2 compares ToLB with state-of-the-art methods on OV-COCO. With ViT-B/16, ToLB achieves 42.5 Novel AP_{50}, improving the CLIPSelf baseline by 4.9 points and surpassing DST-Det by 1.2 points. The improvement remains consistent with ViT-L/14, where ToLB reaches 48.2 Novel AP_{50}, outperforming the baseline by 3.9 points and the previous best result by 1.5 points. Compared with DeCLIP (Wang et al., 2025), which constrains patch-level correlations within CLIP to improve local visual representations, ToLB instead operates at the detector level by preserving inter-region structure after backbone freezing. ToLB outperforms DeCLIP by 1.4 and 2.0 Novel AP_{50} with ViT-B/16 and ViT-L/14, respectively, highlighting the importance of explicitly addressing detector-level base bias.

Table 2: Comparison with state-of-the-art open-vocabulary object detection methods on the OV-COCO benchmark. The best results are marked in bold.

Method	Backbone	Training Data	Novel AP_{50}	Base AP_{50}	All AP_{50}
VLDet (Lin et al., 2023)	RN50	COCO-Base + COCO Cap.	32.0	50.6	45.8
DVDet (Jin et al., 2024)	RN50	COCO-Base + COCO Cap.	34.6	52.8	48.0
RALF (Kim et al., 2024)	RN50	COCO-Base + COCO Cap.	33.4	54.5	49.0
CADet (Liang & Zhang, 2025)	RN50	COCO-Base + COCO Cap.	36.4	50.6	46.9
TLDet (Liang & Zhang, 2026b)	RN50	COCO-Base + COCO Cap.	36.1	54.5	49.7
F-ViT (Kuo et al., 2023)	ViT-B/16	COCO-Base	17.5	41.0	34.9
DST-Det (Xu et al., 2025)	ViT-B/16	COCO-Base	41.3	59.6	54.8
ATAS (Yeo et al., 2025)	ViT-B/16	COCO-Base	37.2	54.8	50.2
DenseVLM (Li et al., 2025)	ViT-B/16	COCO-Base	33.1	52.5	47.4
DeCLIP (Wang et al., 2025)	ViT-B/16	COCO-Base	41.1	57.8	53.5
NoOVD (Zhang et al., 2026)	ViT-B/16	COCO-Base	37.9	56.5	51.6
Baseline (Wu et al., 2024)	ViT-B/16	COCO-Base	37.6	54.9	50.4
ToLB (Ours)	ViT-B/16	COCO-Base	**42.5** (+4.9)	55.5	52.1
F-ViT (Kuo et al., 2023)	ViT-L/14	COCO-Base	24.7	53.6	46.0
DST-Det (Xu et al., 2025)	ViT-L/14	COCO-Base	46.7	61.9	58.0
DeCLIP (Wang et al., 2025)	ViT-L/14	COCO-Base	46.2	65.2	60.3
LoCLIP (Lee et al., 2026)	ViT-L/14	COCO-Base	44.0	65.8	60.0
NoOVD (Zhang et al., 2026)	ViT-L/14	COCO-Base	45.4	64.7	59.7
Baseline (Wu et al., 2024)	ViT-L/14	COCO-Base	44.3	64.1	59.0
ToLB (Ours)	ViT-L/14	COCO-Base	**48.2** (+3.9)	64.1	59.9

Table 3: Comparison with state-of-the-art OVOD methods on OV-LVIS benchmark.

Method	Backbone	Training Data	mAP_r^{mask}	mAP_c^{mask}	mAP_f^{mask}	mAP_{all}^{mask}
RegionCLIP (Zhong et al., 2022)	RN50	LVIS-Base + CC3M	17.1	27.4	34.0	28.2
VLDet (Lin et al., 2023)	RN50	LVIS-Base + CC3M	21.7	29.8	34.3	30.1
RALF (Kim et al., 2024)	RN50	LVIS-Base + CC3M	21.9	26.2	29.1	26.6
F-ViT (Kuo et al., 2023)	ViT-B/16	LVIS-Base	11.5	12.3	20.6	15.4
DST-Det (Xu et al., 2025)	ViT-B/16	LVIS-Base	26.2	-	-	-
DenseVLM (Li et al., 2025)	ViT-B/16	LVIS-Base	23.9	18.4	–	21.4
ATAS (Yeo et al., 2025)	ViT-B/16	LVIS-Base	25.8	21.4	28.9	25.8
DeCLIP (Wang et al., 2025)	ViT-B/16	LVIS-Base	26.8	22.4	29.8	26.0
Baseline (Wu et al., 2024)	ViT-B/16	LVIS-Base	25.3	21.8	29.1	25.2
ToLB (Ours)	ViT-B/16	LVIS-Base	**27.6** (+2.3)	22.9	28.7	26.1
F-ViT (Kuo et al., 2023)	ViT-L/14	LVIS-Base	24.2	27.9	31.5	28.7
DeCLIP (Wang et al., 2025)	ViT-L/14	LVIS-Base	37.2	35.2	36.5	36.0
Baseline (Wu et al., 2024)	ViT-L/14	LVIS-Base	34.9	34.6	35.6	35.1
ToLB (Ours)	ViT-L/14	LVIS-Base	**39.5** (+4.6)	36.0	35.3	36.4

Results on OV-LVIS. Table 3 shows that ToLB consistently improves rare-category recognition on OV-LVIS. With ViT-B/16, ToLB achieves 27.6 mAP_r^{mask}, outperforming the CLIPSelf baseline by 2.3 points and DeCLIP by 0.8 points, while also improving the overall mAP^{mask} from 25.2 to 26.1. With ViT-L/14, ToLB further reaches 39.5 mAP_r^{mask}, yielding gains of 4.6 points over the baseline and 2.3 points over DeCLIP. These consistent improvements on OV-LVIS further demonstrate the effectiveness and generalization of our proposed ToLB in OVOD.

Component ablation. As shown in Table 4, both RKP and VRR consistently improve novel-category detection. RKP increases Novel AP_{50} from 37.6 to 41.0 (+3.4), which validates the effectiveness of relational preservation in mitigating representation bias. VRR independently improves Novel AP_{50} to 39.4 (+1.8), confirming the benefit of category-agnostic spatial refinement. Combining the two modules achieves 42.5 Novel AP_{50} (+4.9), while Base AP_{50} remains nearly unchanged. This complementary improvement supports our design of addressing representation and localization bias at training and inference, respectively.

Choice of structural teacher. We further study the choice of the teacher model for RKP in Table 5. Using CLIP as the teacher improves Novel AP_{50} from 37.6 to 38.9, showing that relational supervi-

Table 4: Ablation study of the proposed components on OV-COCO.

Method	Novel AP_{50}	Base AP_{50}	All AP_{50}
CLIPSelf	37.6	54.9	50.4
+ RKP	41.0	55.6	51.8
+ VRR	39.4	55.4	51.2
ToLB	**42.5**	55.5	52.1

Table 5: Ablation study of the structural teacher used in RKP on OV-COCO.

Method	Novel AP_{50}	Base AP_{50}	All AP_{50}
CLIPSelf	37.6	54.9	50.4
RKP *w.* CLIP	38.9	55.2	50.9
RKP *w.* DINOv2	**41.0**	55.6	51.8

sion itself is beneficial. Replacing CLIP with DINOv2 further increases Novel AP_{50} to 41.0, yielding an additional gain of 2.1 points while also improving Base and All AP_{50}. This result supports our choice of using DINOv2 as the structural teacher, as its category-agnostic visual representations provide a more suitable prior for preserving inter-region structure than CLIP features.

Ablation on knowledge preservation strategies. To examine the effectiveness of different knowledge preservation strategies, we compare several alternatives in Table 6. Direct feature alignment reduces Novel AP_{50} from 37.6 to 32.3, suggesting that absolute matching across heterogeneous feature spaces can harm open-vocabulary transfer. Relational preservation is more effective, but applying it to all regions yields only marginal gains. Restricting preservation to detector-background regions improves performance, while mining potential novel regions further increases Novel AP_{50} to 41.0, outperforming random background sampling by 1.8 points and the baseline by 3.4 points. These results validate both relational preservation and targeted region selection in RKP.

Table 6: Ablation study on different knowledge preservation strategies for RKP on OV-COCO.

Preservation Strategy	Novel AP_{50}
Baseline	37.6
Direct Feature Alignment	32.3
Relation on All Regions	37.8
Relation on Random BG Regions	39.2
RKP on Potential Regions	**41.0**

4.3 Qualitative Analysis

Qualitative comparison. Qualitative comparisons with CLIPSelf on novel categories are provided in the Appendix A.6, illustrating the improved detection ability for novel objects. We further analyze novel-to-base misclassifications in Appendix A.5. Our ToLB reduces the misclassification rate from 25.04% to 13.24%, demonstrating alleviated base-category bias towards novel objects.

Qualitative analysis of VRR. Figure 5 qualitatively illustrates the effect of VRR on proposal localization. Without VRR, the baseline often produces loose regions containing substantial background context. By exploiting category-agnostic VLM objectness cues, VRR suppresses irrelevant background regions and produces more object-centric proposals. This effect is consistently observed across different object categories and scenes, supporting the role of VRR in alleviating spatial bias for novel objects.

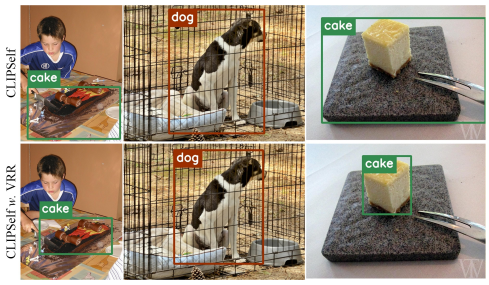

Figure 5: Qualitative effect of VRR.

5 Conclusion

In this paper, we identify that freezing the visual backbone is insufficient to eliminate detector-level base bias, and base-only optimization distorts novel-region representations, limiting the transferability of pretrained visual knowledge. To overcome these issues, we introduce ToLB, which mitigates representation and localization bias from complementary training and inference-time perspectives. Experiments on OV-COCO and OV-LVIS consistently improve novel-category performance across different backbone scales while largely preserving base-category recognition. These results suggest that effective open-vocabulary transfer requires not only preserving pretrained representations, but also controlling how downstream detector components reshape them during task-specific adaptation.

REFERENCES

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. URL https://arxiv.org/abs/1504.00325.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5356–5364, 2019.

Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. In *International Conference on Learning Representations*, 2024.

Jooyeon Kim, Eulrang Cho, Sehyung Kim, and Hyunwoo J Kim. Retrieval-augmented open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17427–17436, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations*, 2023.

Sanghoon Lee, Geon Lee, Hyekang Park, and Bumsub Ham. Exploring hierarchical consistency and unbiased objectness for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 6819–6828, 2026.

Yunheng Li, Yuxuan Li, Quan-Sheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased region-language alignment for open-vocabulary dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23795–23805, 2025.

Qiuyu Liang and Yongqiang Zhang. Sam based region-word clustering and inference score adjusting for open-vocabulary object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2596–2605, 2025.

Qiuyu Liang and Yongqiang Zhang. Learning local and consistent semantics for open-vocabulary dense prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026a.

Qiuyu Liang and Yongqiang Zhang. Tensor decomposition and language description for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6853–6861, 2026b.

Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *International Conference on Learning Representations*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:71078–71094, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, pp. 1–31, 2024.

Chau Pham, Truong Vu, and Khoi Nguyen. Lp-ovod: Open-vocabulary object detection by linear probing. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 779–788. IEEE, 2024.

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 *Int. Conf. Machine. Learning*, pp. 8748–8763. PMLR, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *The Association for Computational Linguistics*, pp. 2556–2565, 2018.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14824–14834, 2025.

Kuo Wang, Lechao Cheng, Weikai Chen, Pingping Zhang, Liang Lin, Fan Zhou, and Guanbin Li. Marvelovd: Marrying object recognition and vision-language models for robust open-vocabulary object detection. In *European Conference on Computer Vision*, pp. 106–122. Springer, 2024.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15254–15264, 2023.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *International Conference on Learning Representations*, 2024.

Shilin Xu, Xiangtai Li, Size Wu, Wenwei Zhang, Yunhai Tong, and Chen Change Loy. Dst-det: Open-vocabulary object detection via dynamic self-training. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):5037–5050, 2025. doi: 10.1109/TCSVT.2024.3520734.

Juan Yeo, Soonwoo Cha, Jiwoo Song, Hyunbin Jin, and Taesup Kim. Atas: Any-to-any self-distillation for enhanced open-vocabulary dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20390–20400, 2025.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, 2021.

Chuhan Zhang, Chaoyang Zhu, Pingcheng Dong, Long Chen, and Dong Zhang. Cyclic Contrastive Knowledge Transfer for Open-Vocabulary Object Detection. In *International Conference on Learning Representations*, 2025. URL `https://mlanthology.org/iclr/2025/zhang2025iclr-cyclic/`.

Yupeng Zhang, Ruize Han, Zhiwei Chen, Wei Feng, and Liang Wan. Noovd: Novel category discovery and embedding for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6304–6313, June 2026.

Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European conference on computer vision*, pp. 159–175. Springer, 2022.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

AI USE STATEMENT

Generative AI tools were used for code assistance and manuscript editing to improve clarity and readability. They were not used to formulate the core research ideas or conclusions. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final work.

ETHICS STATEMENT

This work uses publicly available datasets and does not involve human subjects, private data, or sensitive personal information.

REPRODUCIBILITY STATEMENT

We provide detailed implementation settings, training configurations, and hyper-parameters in the main paper.

A APPENDIX

In this appendix, we provide additional implementation details and analyzes for ToLB, including the pseudocode of RKP and VRR, hyperparameter ablations, diagnostic analysis of novel-to-base misclassifications, and qualitative results on novel categories. Specifically:

- Appendix A.1 presents the pseudocode of Relational Knowledge Preservation (RKP), including potential novel-region mining and relational structure preservation.
- Appendix A.2 presents the pseudocode of VLM-Guided Region Refinement (VRR), detailing the category-agnostic objectness-based proposal refinement procedure.
- Appendix A.3 studies the effect of the trade-off parameter λ_{RKP}.
- Appendix A.4 analyzes the Top-K retention ratio used in VRR.
- Appendix A.5 investigates novel-to-base misclassifications to further analyze base-category bias.
- Appendix A.6 provides additional qualitative comparisons on novel categories.

All ablation experiments are conducted with the same ViT-B/16 backbone unless otherwise specified.

A.1 PSEUDOCODE OF RELATIONAL KNOWLEDGE PRESERVATION

Algorithm 1 summarizes the training-time procedure of RKP, including potential novel-region mining and relational structure preservation.

A.2 PSEUDOCODE OF VLM-GUIDED REGION REFINEMENT

Algorithm 2 summarizes the inference-time procedure of VRR, which refines RPN proposals using category-agnostic VLM objectness cues.

A.3 ABLATION STUDY ON THE TRADE-OFF PARAMETER

Table 7: Ablation study on λ_{RKP}.

λ_{RKP}	Novel AP_{50}	**Base AP_{50}**	**All AP_{50}**
1.0	41.7	55.4	51.8
3.0	42.3	55.4	51.9
5.0	**42.5**	55.5	52.1
7.0	42.1	55.3	51.8
9.0	41.6	55.3	51.7

We analyze the sensitivity of the trade-off parameter λ_{RKP} in Eq. 9, which controls the strength of relational preservation. As shown in Table 7, Novel AP_{50} increases from 41.7 to 42.5 as λ_{RKP} increases from 1.0 to 5.0, and gradually decreases when the weight is further increased. The best performance is achieved at $\lambda_{\mathrm{RKP}} = 5.0$, suggesting that overly weak regularization provides insufficient structural preservation, whereas an excessively strong constraint may restrict detector adaptation. We therefore set $\lambda_{\mathrm{RKP}} = 5.0$ for all experiments.

Algorithm 1 Relational Knowledge Preservation (RKP)

Require: Image I, sampled RoIs $\mathcal{R}$ with labels $\{c_i\}$, frozen VLM Φ_{VLM}, frozen VFM Φ_{VFM}, detector Φ_{Det}
Ensure: $\mathcal{L}_{\text{RKP}}$
$\mathcal{R}_{bg} \leftarrow \{r_i \in \mathcal{R} \mid c_i = c_{bg}\}$
Encode generic foreground/background prompts as $\mathbf{t}_{fg}$ and $\mathbf{t}_{bg}$
$\mathcal{R}_{pn} \leftarrow \emptyset$
for $r_i \in \mathcal{R}_{bg}$ **do**
    $\mathbf{z}_i \leftarrow \Phi^{V}_{\text{VLM}}(I, r_i)$
    $s_i^{fg} \leftarrow \cos(\mathbf{z}_i, \mathbf{t}_{fg})$, $s_i^{bg} \leftarrow \cos(\mathbf{z}_i, \mathbf{t}_{bg})$
    if $s_i^{fg} > s_i^{bg}$ **then**
        $\mathcal{R}_{pn} \leftarrow \mathcal{R}_{pn} \cup \{r_i\}$
    end if
end for
for $r_i \in \mathcal{R}_{pn}$ **do**
    $\mathbf{v}_i \leftarrow \Phi_{\text{VFM}}(I, r_i)$
    $\mathbf{h}_i \leftarrow \Phi_{\text{Det}}(I, r_i)$
end for
$\mathbf{A}^{\text{VFM}} \leftarrow \text{PairwiseCosine}(\{\mathbf{v}_i\})$
$\mathbf{A}^{\text{Det}} \leftarrow \text{PairwiseCosine}(\{\mathbf{h}_i\})$
Set $\text{diag}(\mathbf{A}^{\text{VFM}}) = \text{diag}(\mathbf{A}^{\text{Det}}) = 0$
$\mathcal{L}_{\text{RKP}} \leftarrow \text{SmoothL1}(\mathbf{A}^{\text{Det}}, \mathbf{A}^{\text{VFM}})$
return $\mathcal{L}_{\text{RKP}}$

Algorithm 2 VLM-Guided Region Refinement (VRR)

Require: Image I, RPN proposals $\mathcal{R}$, frozen VLM $(\Phi_{\text{V}}, \Phi_{\text{T}})$, retention ratio ρ
Ensure: Refined proposals $\widehat{\mathcal{R}}$
$\mathbf{t}_{obj} \leftarrow \Phi_{\text{T}}(\text{“a photo of an object”})$
$\mathbf{F} \leftarrow \Phi_{\text{V}}(I)$
$\mathbf{O} \leftarrow \text{CosineMap}(\mathbf{F}, \mathbf{t}_{obj})$
$\widehat{\mathcal{R}} \leftarrow \emptyset$
for $r_i \in \mathcal{R}$ **do**
    $\mathbf{S}_i \leftarrow \text{RoIAlign}(\mathbf{O}, r_i)$
    $K \leftarrow \lceil \rho R^2 \rceil$
    $\Omega_i \leftarrow \text{TopK}(\mathbf{S}_i, K)$
    $\hat{r}_i \leftarrow \text{MapToImage}\left(\text{BBox}(\Omega_i), r_i\right)$
    $\widehat{\mathcal{R}} \leftarrow \widehat{\mathcal{R}} \cup \{\hat{r}_i\}$
end for
return $\widehat{\mathcal{R}}$

A.4 Ablation study on the Top-K parameter

To analyze the effect of foreground selection in VRR, we vary the Top-K retention ratio in Eq. 7 of the main paper, with the results reported in Table 8. Retaining only 20% yields 41.3 Novel AP_{50}, suggesting that overly aggressive selection may discard useful object regions. Performance peaks at 40% with 42.5 Novel AP_{50}, while larger ratios remain stable at 42.2–42.3. This indicates that a moderate retention ratio provides a better balance between preserving object extent and suppressing background responses. Meanwhile, Base AP_{50} remains nearly unchanged across all settings, showing that VRR is robust to this hyperparameter.

Table 8: Ablation study on the Top-K parameter.

Top-K (%)	Novel AP_{50}	**Base AP_{50}**	**All AP_{50}**
20	41.3	55.5	51.8
40	**42.5**	55.5	52.1
60	42.3	55.4	52.0
80	42.2	55.5	52.0
100	42.2	55.4	51.9



Figure 6: The qualitative comparisons on OV-COCO novel categories. The top row shows predictions from the CLIPSelf (Wu et al., 2024) while the bottom row shows our results. Our method improves novel-object recall and localization, reduces duplicate or incomplete predictions, and recovers several novel objects missed by the baseline, including *bus*, *sink*, and *cat*.

A.5 Analysis of Novel-to-Base Misclassifications

We further investigate whether our method alleviates the tendency of novel objects to be incorrectly absorbed by base categories. Specifically, for each novel ground-truth instance, we identify its matched prediction with $\text{IoU} > 0.5$ and measure whether the predicted category belongs to the base-category set. The resulting novel-to-base misclassification rate directly reflects the extent to which novel objects are semantically biased toward seen categories.

As shown in Table 9, the baseline yields an overall misclassification rate of 25.04%, whereas our method substantially reduces it to 13.24%. This corresponds to an absolute reduction of 11.80 percentage points and a relative reduction of 47.1%. The improvement is also consistent across nearly all novel categories. In particular, large reductions are observed for *cup* ($37.88\% \rightarrow 17.17\%$), *couch* ($56.32\% \rightarrow 37.69\%$), *knife* ($30.77\% \rightarrow 15.69\%$), and *cat* ($37.31\% \rightarrow 23.16\%$). These results provide further evidence that our method mitigates the semantic bias induced by base-only detector optimization, making novel regions less likely to drift toward base-category representations.

Table 9: Novel-to-base misclassification analysis on OV-COCO. We report the percentage of novel ground-truth instances whose matched predictions ($\text{IoU} > 0.5$) are assigned to base categories. Lower is better. Δ denotes the absolute reduction over the baseline.

Novel Category	Baseline (%)	Ours (%)	Δ
airplane	12.41	**7.96**	4.45
bus	22.26	**10.95**	11.31
cake	29.68	**18.12**	11.56
cat	37.31	**23.16**	14.15
couch	56.32	**37.69**	18.63
cow	22.04	**9.68**	12.36
cup	37.88	**17.17**	20.71
dog	29.82	**19.62**	10.20
elephant	6.40	**0.40**	6.00
keyboard	13.73	**7.84**	5.89
knife	30.77	**15.69**	15.08
scissors	16.67	**13.89**	2.78
sink	23.32	**10.45**	12.87
skateboard	6.15	**3.35**	2.80
snowboard	43.48	**31.88**	11.60
tie	0.40	**0.40**	0.00
umbrella	6.88	**2.46**	4.42
Overall	25.04	**13.24**	**11.80**

A.6 Qualitative Results on Novel Categories

Figure 6 compares our method with the baseline (Wu et al., 2024) on novel categories. Our method reduces duplicate or incomplete detections and improves both recall and localization, successfully recovering missed objects such as *bus*, *sink*, and *cat*, while producing more complete boxes for *umbrella* and additional instances such as *dog*. These results qualitatively support that our method better preserves novel-object semantics and produces more object-centric regions, consistent with the novel-to-base analysis in Appendix A.5.