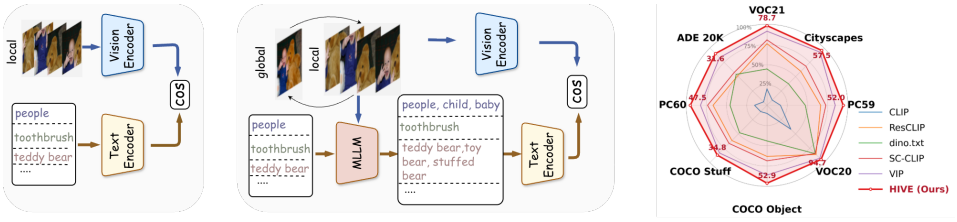
HIVE: HOLISTIC IMAGE-ADAPTIVE VOCABULARY EVOLUTION FOR TRAINING-FREE DENSE VISION-LANGUAGE INFERENCE

Anonymous authors
Paper under double-blind review



(a) Conventional inference paradigm

(b) Inference paradigm of HIVE

(c) Performance comparison with state-of-the-arts

Figure 1: Paradigm and performance comparison between conventional approach and our HIVE.

ABSTRACT

Benefiting from pretrained vision-language models, training-free open-vocabulary semantic segmentation (OVSS) predicts pixel-level labels for arbitrary categories without task-specific optimization. Existing training-free OVSS methods commonly perform dense prediction using independently processed local views and a fixed set of category names. However, this paradigm suffers from two limitations. First, processing local views independently prevents the model from capturing cross-view dependencies, limiting contextual reasoning for dense prediction. Second, fixed category names describe semantic classes generically and fail to cover the image-specific appearances of target objects, weakening visual-text alignment. To overcome these limitations, we propose **H**olistic **I**mage-adaptive **V**ocabulary **E**volution (**HIVE**), a training-free framework that jointly improves visual representations and textual vocabularies. Specifically, HIVE introduces Holistic Visual Reasoning (HVR) yielding holistic representations through local-global-local interaction. Furthermore, Image-adaptive Vocabulary Evolution (IVE) constructs a dedicated vocabulary set for each image. It first employs an off-the-shelf MLLM to generate image-conditioned category aliases and then iteratively evolutes the vocabulary set by selecting beneficial aliases under semantic and structural visual guidance. Experiments across eight evaluation protocols show that HIVE achieves an average mIoU of 56.2%, outperforming existing training-free OVSS methods and establishing new state-of-the-art performance. Our code will be publicly available after acceptance.

1 INTRODUCTION

Open-vocabulary semantic segmentation (OVSS) aims to assign pixel-level labels from an arbitrary given category set, allowing segmentation models to recognize categories beyond a fixed training taxonomy. Pretrained vision-language models (VLMs), such as CLIP (Radford et al., 2021), provide a natural foundation for OVSS by aligning visual representations with free-form text queries. Inspired by this, recent methods (Zhou et al., 2022; Wang et al., 2024; Lan et al., 2024a;b; Hajimiri et al., 2025; Yang et al., 2025; Chi et al., 2025) repurpose frozen VLMs for dense prediction without task-specific optimization or mask annotations, dubbed as training-free OVSS.

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.

The prevailing paradigm for training-free OVSS can be illustrated in Fig. 1(a), where dense predictions are obtained from independently processed local views using a fixed vocabulary. However, this paradigm suffers from two limitations. First, processing these local views independently prevents the model from capturing cross-view dependencies, limiting contextual reasoning for dense prediction (Zhou et al., 2022; Lan et al., 2024a;b; Hajimiri et al., 2025). Second, a single category name may not cover the diverse visual forms appearing in a particular image (Lan et al., 2024b; Shi et al., 2025). For example, the category name `people` for the character in image is coarse.

To address these limitations, we propose **H**olistic **I**mage-adaptive **V**ocabulary **E**volution (**HIVE**), a training-free OVSS framework that jointly improves visual representations and textual vocabularies. As illustrated in Fig. 1(b), HIVE consists of Holistic Visual Reasoning (HVR) and Image-adaptive Vocabulary Evolution (IVE). HVR enhances local representations through local-global-local interaction. To reduce the potential semantic bias introduced by the coarse global view, HVR first refines the global representation with fine-grained local cues and then uses it as a bridge to propagate context across local views. IVE constructs an image-specific vocabulary set by first employing an MLLM to generate candidate aliases according to the visual content of each image. It then iteratively evolves the vocabulary by selecting beneficial candidates based on local-global semantic consistency and visual structural consistency. Finally, HIVE matches the refined local representations with the image-specific vocabulary and applies affinity-guided random walks to obtain the final segmentation. Compared with GLA-CLIP (Lee et al., 2026), HIVE introduces an explicit global view to connect isolated local views with whole-image context, while refining its coarse representation with fine-grained local cues before context propagation. Meanwhile, HIVE improves upon the image-shared vocabulary (Zhu et al., 2026) by constructing a dedicated vocabulary for each image, thereby reducing irrelevant semantics and better matching its visual content. We evaluate our method on eight popular training-free OVSS dataset protocols, including PASCAL VOC (w/ and w/o background), PASCAL Context (w/ and w/o background), COCO-Object, COCO-Stuff164k, Cityscapes and ADE20K. The experimental results demonstrates its state-of-the-art performance with an average mIoU of 56.2% across eight protocols, outperforming previous work by 3.8%.

Our main contributions can be summarized as follows:

- We propose a Holistic Image-adaptive Vocabulary Evolution (HIVE) framework for training-free OVSS. By coupling holistic cross-view reasoning with image-adaptive vocabulary evolution, HIVE improves visual-text alignment without task-specific training.
- We devise a Holistic Visual Reasoning (HVR) module to enhance fine-grained local representations with long-range context. By introducing an explicit global view, HVR performs local-global-local interaction to capture cross-view dependencies for dense prediction.
- We develop an Image-adaptive Vocabulary Evolution (IVE) module to construct a dedicated vocabulary set for each image. By augmenting category names with MLLM-generated image-conditioned aliases and iteratively selecting candidates under visual guidance, IVE yields more relevant and discriminative text words, improving visual-text alignment.
- Extensive experiments across eight OVSS evaluation protocols demonstrate that HIVE achieves state-of-the-art performance, while comprehensive ablations validate the effectiveness of each component.

2 Related Work

2.1 Training-Free Open-Vocabulary Semantic Segmentation

To reduce the supervision and optimization costs of training-based methods (Li et al., 2022; Xu et al., 2022; Ghiasi et al., 2022; Cho et al., 2024), training-free open-vocabulary semantic segmentation (OVSS) repurposes frozen VLMs for dense prediction. Early studies improve CLIP's spatial sensitivity through attention reconstruction (Wang et al., 2024) or residual removal (Lan et al., 2024a). Recent methods incorporate spatial priors from vision foundation models: ProxyCLIP (Lan et al., 2024b) replaces CLIP attention with DINO affinities, Trident (Shi et al., 2025) aggregates stitched crop features using global SAM correlations (Kirillov et al., 2023), and dino.txt (Jose et al., 2025) aligns language with frozen DINO features. High-resolution inference, however, often processes

windows independently, with cross-window context introduced at the final attention layer (Lee et al., 2026) or after feature stitching (Moon et al., 2026). HIVE instead couples an explicit global view with local windows through closed-loop feature and relation interactions, producing holistic features for dense inference and affinities for vocabulary evolution.

2.2 Text Prompt Refinement

LLM-generated class descriptions improve text representations for zero-shot learning (Menon & Vondrick, 2023; Sun et al., 2024). VIP (Zhu et al., 2026) expands class names into LLM-generated aliases and selects them using DINO structural evidence, but its shared vocabulary may introduce irrelevant semantics and weaken visual-text alignment. Image-adaptive methods include FreeCP (Chen et al., 2025), which filters redundant or ambiguous categories using activation structure; FSeg-LLaVA (Zhang et al., 2026), which queries an MLLM for class presence and extracts segmentation cues from its internal responses; and GR-CoT (Zhou et al., 2026), which uses multimodal reasoning to construct candidate vocabularies for remote-sensing segmentation. However, these methods perform dense vision-language inference on pruned candidate categories, suffering from the risk of missing valid categories. In contrast, HIVE retains all original categories, generates aliases only for MLLM-identified classes, and greedily evaluates the entire vocabulary set using categorical and structural visual consistency to construct an image-specific vocabulary set.

3 Methodology

3.1 Preliminaries

Dense Vision-Language Inference. Given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a category set $\mathcal{C} = \{c_1, \ldots, c_C\}$, the goal is to predict a pixel-wise label map $\mathbf{Y} \in \mathbb{R}^{H\times W}$ without parameter updates or mask annotations. The image is divided into R overlapping local views $\mathbf{I}_{L_r}{}_{r=1}^{R}$. For each local view, patch-level logits $\mathbf{Z}_{L_r} \in \mathbb{R}^{N_r\times C}$ are obtained by matching its spatial representations $\mathbf{F}_{L_r} \in \mathbb{R}^{N_r\times d}$ with the textual category embeddings $\mathbf{T} \in \mathbb{R}^{C\times d}$. The local logits are then upsampled, mapped to their corresponding locations on the image canvas, and fused to produce the image-level prediction scores $\mathbf{P} \in \mathbb{R}^{H\times W\times C}$. The prediction is formulated as:

$$\mathbf{Z}_{L_r} = \mathrm{cosine}(\mathbf{F}_{L_r}, \mathbf{T}), \quad \mathbf{P} = \mathrm{Softmax}\left(\mathcal{F}_{\mathrm{logit}}\left(\mathcal{U}(\mathbf{Z}_{L_r})_{r=1}^{R}\right)\right), \quad \mathbf{Y}_u = \arg\max_{c\in\mathcal{C}} \mathbf{P}_{u,c}. \tag{1}$$

Here, $\mathcal{U}$ upsamples each local logit map and places it at the corresponding location on the image canvas, while $\mathcal{F}_{\mathrm{logit}}$ fuses the aligned local logits by averaging their overlapping regions.

Self-Corrected Vision Encoder Baseline. To enhance spatial awareness for OVSS, we follow previous work (Zhu et al., 2026) and employ a frozen DINO backbone (Siméoni et al., 2025) as the vision encoder $\mathscr{E}$. HVR builds on a lightweight text-alignment adapter $\mathscr{A}$. To preserve the spatial structure encoded by DINO, we follow ProxyCLIP (Lan et al., 2024b) and replace the adapter's native attention weights with proxy attention derived from the self-similarity of DINO patch tokens $\mathbf{X} \in \mathbb{R}^{N\times d}$. This involves affinity calibration, negative-affinity masking, and normalization:

$$\mathbf{E} = \gamma\left(\mathbf{X}\mathbf{X}^\top - \frac{\beta}{N^2}\sum_{i,j}[\mathbf{X}\mathbf{X}^\top]_{i,j}\right), \quad \mathbf{M}_{i,j} = \begin{cases} 0, & \mathbf{E}_{i,j} \geq 0, \\ -\infty, & \mathbf{E}_{i,j} < 0, \end{cases} \quad \mathbf{G} = \mathrm{Softmax}\left(\mathbf{E} + \mathbf{M}\right). \tag{2}$$

Here, β and γ control the affinity shift and scale, respectively. For brevity, we denote this complete operation as $\mathbf{G} = \mathrm{Softmax}_+(\mathbf{X}\mathbf{X}^\top)$ in subsequent formulations.

3.2 Overview

In the conventional paradigm described in Eqn. 1, dense predictions are obtained from independently processed local views using a fixed vocabulary. However, this paradigm neglects cross-view visual dependencies and image-specific linguistic variations, limiting the discriminability of vision-language representations. To address these limitations, we incorporate cross-view context to obtain fine-grained local visual descriptors and construct an image-adaptive vocabulary set for each input image.

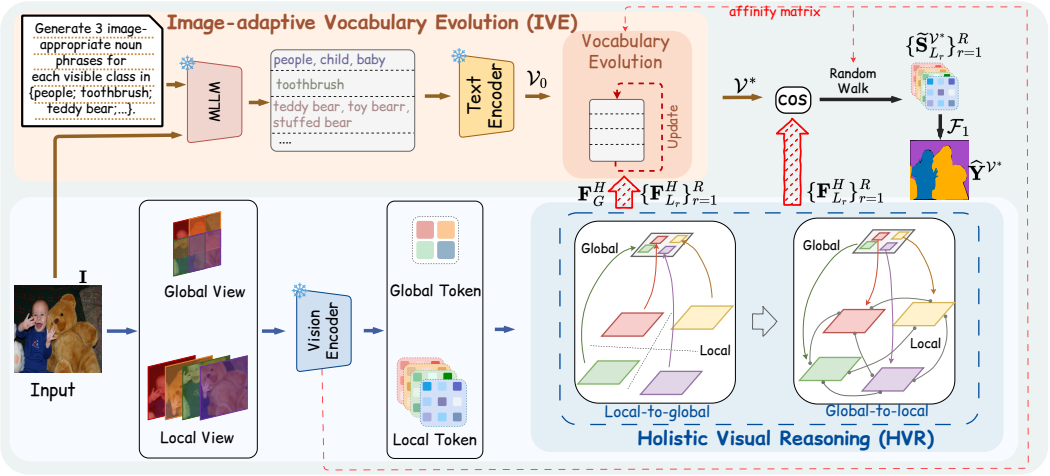


Figure 2: Overview of HIVE. A frozen visual encoder processes a global view and multiple overlapping local views. HVR refines local features through local-to-global-to-local interaction, producing holistic local representations $\{\mathbf{F}^H_{L_r}\}_{r=1}^R$. IVE iteratively evolves an image-specific vocabulary set $\mathcal{V}^*$ under visual guidance. Finally, the local predictions derived from $\{\mathbf{Z}^{\mathcal{V}^*}_{L_r}\}_{r=1}^R$ are refined and fused to produce the final segmentation $\widetilde{\mathbf{Y}}^{\mathcal{V}^*}$.

As illustrated in Fig. 2, we construct a multi-view visual input comprising a global view and R overlapping local views, all resized to the same input resolution and independently processed by a frozen visual encoder $\mathscr{E}$. The resulting multi-view tokens are fed into a Holistic Visual Reasoning (HVR) module, which models global-local interactions and produces a holistic visual representation $\mathbf{F}^H_{L_r}$ for each local view L_r. Based on the holistic local representations and local structural affinities, our Image-adaptive Vocabulary Evolution (IVE) module constructs an image-specific vocabulary set $\mathcal{V}^*$.

For the vocabulary set $\mathcal{V}^*$, we compute class logits $\mathbf{Z}^{\mathcal{V}^*}_{L_r}$ by matching $\mathbf{F}^H_{L_r}$ with the corresponding text embeddings. We then apply argmax and one-hot encoding to obtain the local prediction map $\mathbf{S}^{\mathcal{V}^*}_{L_r}$. To encourage consistent predictions among structurally related patches, we refine these maps through a κ-step random walk (Ahn & Kwak, 2018) guided by the local visual structure:

$$\widetilde{\mathbf{S}}^{\mathcal{V}^*}_{L_r} = \mathrm{Norm}\left((\mathbf{A}_{L_r})^{\kappa}\,\mathbf{S}^{\mathcal{V}^*}_{L_r}\right), \quad \widetilde{\mathbf{Y}}^{\mathcal{V}^*} = \mathcal{F}_{\mathrm{pred}}\left(\{\widetilde{\mathbf{S}}^{\mathcal{V}^*}_{L_r}\}_{r=1}^R\right). \tag{3}$$

Here, the transition matrix $\mathbf{A}_{L_r}$ is obtained by averaging the attention map from the all visual encoder layers for local view r. It captures the structural relations among patches and guides prediction propagation within local view L_r. Norm applies min-max normalization independently to each class map within each local view. The operator $\mathcal{F}_{\mathrm{pred}}$ upsamples and fuses the local maps, averages overlapping regions, and applies pixel-wise argmax. The resulting map $\widetilde{\mathbf{Y}}^{\mathcal{V}^*}$ serves as the final segmentation output.

3.3 Holistic Visual Reasoning (HVR)

Local views preserve fine-grained details but are processed independently, preventing them from capturing long-range contextual dependencies across views. Although the global view provides long-range context, its coarse representation loses fine-grained visual cues. To address these limitations, HVR introduces a local-to-global-to-local interaction. Fine-grained local cues first refine the corresponding regions of the global representation (local-to-global), which then serves as a shared bridge to propagate cross-view context back to each local view (global-to-local). We implement this interaction by modifying the attention layers of the frozen adapter $\mathscr{A}$.

Let $\mathbf{X}_G \in \mathbb{R}^{N_G \times d}$ and $\mathbf{X}_{L_r} \in \mathbb{R}^{N_r \times d}$ denote the global and r-th local tokens extracted by the vision encoder $\mathscr{E}$, respectively. The original value projections in the current attention layer of $\mathscr{A}$ are denoted as $\mathbf{V}_G$ and $\mathbf{V}_{L_r}$ for the global and r-th local views, respectively. We first perform

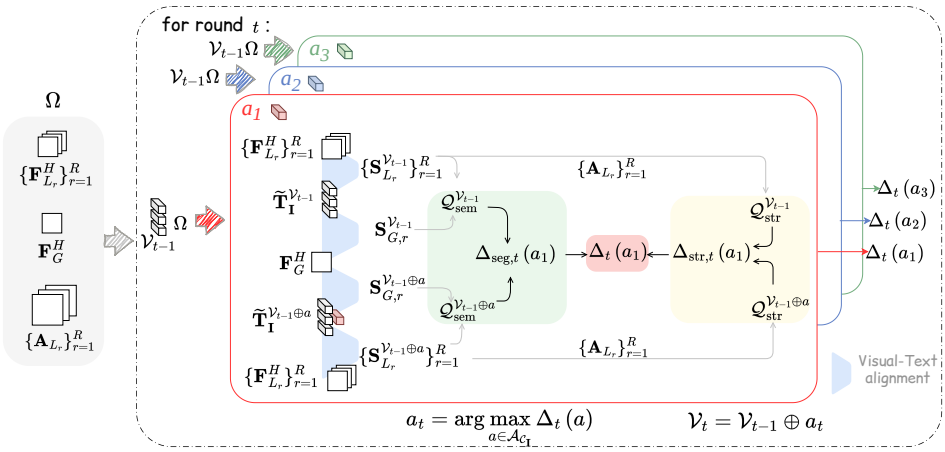


Figure 3: Illustration of one round of vocabulary evolution in IVE. At round t, IVE takes the holistic representations $\{\mathbf{F}^H_{L_r}\}^R_{r=1}$ and $\mathbf{F}^H_G$, the local affinity matrices $\{\mathbf{A}_{L_r}\}^R_{r=1}$, and the current vocabulary set $\mathcal{V}_{t-1}$ as inputs. It evaluates each remaining alias $a \in \mathcal{A}^t_{\mathcal{C}_\mathbf{I}}$ using its gain $\Delta_t(a)$, selects a_t, and updates the vocabulary set to $\mathcal{V}_t$. Each $\mathbf{A}_{L_r}$ is derived from the outputs of $\mathscr{E}$ for local view L_r.

local-to-global (L2G) interaction to enrich each global region with its corresponding local view:

$$\mathbf{F}^{\text{L2G}}_{L_r} = \text{Softmax}_+\left(\mathbf{X}_{G(\Gamma_r)}[\mathbf{X}_G;\mathbf{X}_{L_r}]^\top\right)[\mathbf{V}_G;\mathbf{V}_{L_r}], \quad \mathbf{F}^{\text{L2G}}_G = \mathcal{F}_{\text{HVR}}\left(\mathbf{F}^{\text{L2G}\,R}_{L_r\quad r=1}\right). \tag{4}$$

Here, $\mathbf{X}_{G(\Gamma_r)}$ denotes the tokens selected from $\mathbf{X}_G$ that geometrically correspond to local view r. The notation $[\cdot;\cdot]$ denotes concatenation along the token dimension. The fusion operator $\mathcal{F}_{\text{HVR}}$ integrates the refined regional messages from all local views, yielding a global representation $\mathbf{F}^{\text{L2G}}_G$ enriched with fine-grained local evidence.

Having enriched the global representation with fine-grained evidence, we further use it as a shared bridge to connect the otherwise isolated local views. Specifically, we perform global-to-local (G2L) interaction to propagate the aggregated context back to each local stream:

$$\mathbf{F}^{\text{G2L}}_{L_r} = \text{Softmax}_+\left(\mathbf{X}_{L_r}[\mathbf{X}_{L_r};\mathbf{X}_G]^\top\right)[\mathbf{V}_{L_r};\mathbf{F}^{\text{L2G}}_G]. \tag{5}$$

The resulting message $\mathbf{F}^{\text{G2L}}_{L_r}$ is subsequently processed by the original output projection, residual connection, and feed-forward network in $\mathscr{A}$. We perform this L2G–G2L interaction independently within each attention layer using its own value projections, and denote the outputs of the final layer as $\{\mathbf{F}^H_{L_r}\}^R_{r=1}$. Here, we denote $\mathbf{F}^{\text{L2G}}_G$ as $\mathbf{F}^H_G$, which is subsequently fed into IVE.

3.4 Image-Adaptive Vocabulary Evolution (IVE)

A category name describes a category generically and may not cover all visual forms appearing in a particular image. To address this limitation, we propose Image-adaptive Vocabulary Evolution (IVE), which constructs a dedicated vocabulary set $\mathcal{V}^*$ for each image based on its visual content. IVE first employs an off-the-shelf MLLM for image-adaptive semantic expansion and then iteratively evolves the vocabulary by jointly considering local-global semantic consistency and visual structural consistency.

Pipeline. We first generate B image-conditioned, class-preserving noun phrases for each visible category $c \in \mathcal{C}_\mathbf{I}$ in image $\mathbf{I}$, yielding the complete alias candidate pool $\mathcal{A}_{\mathcal{C}_\mathbf{I}}$. The visible category set $\mathcal{C}_\mathbf{I}$ is identified by an MLLM. These aliases may include synonyms or descriptions of visually grounded subcategories, attributes, and concrete variants. For example, in Fig. 2, no aliases are generated for `toothbrush`, whose original name remains in the vocabulary set. We generate aliases for visible categories such as {`people; teddy bear`}, yielding the candidate pool $\mathcal{A}_{\mathcal{C}_\mathbf{I}} = \{$`child, baby; toy bear, stuffed bear`$\}$.

Given $\mathcal{A}_{\mathcal{C}_\mathbf{I}}$ and the original category names $\mathcal{C}$, we seek an image-adaptive vocabulary set $\mathcal{V}^*$ for dense classification of image $\mathbf{I}$. We initialize $\mathcal{V}_0 = \mathcal{C}$ and iteratively expand it through greedy

selection. At round t, each remaining candidate alias a is evaluated according to its semantic gain $\Delta_{\text{sem},t}(a)$ and structural consistency gain $\Delta_{\text{str},t}(a)$. Their sum defines the total gain $\Delta_t(a)$, and we select the candidate that maximizes this gain while improving both consistency measures:

$$\Delta_t(a) = \Delta_{\text{sem},t}(a) + \Delta_{\text{str},t}(a), \quad a_t = \underset{a \in \mathcal{A}^t_{\mathcal{C}_\mathbf{I}}, ; \Delta_{\text{sem},t}(a)>0, ; \Delta_{\text{str},t}(a)>0}{\arg\max} \Delta_t(a). \tag{6}$$

Here, $\mathcal{A}^t_{\mathcal{C}_\mathbf{I}}$ denotes the remaining alias candidate pool at round t, and $\oplus$ denotes adding an alias to the vocabulary set. After selecting a_t, we update the vocabulary and candidate pool as $\mathcal{V}_t = \mathcal{V}_{t-1} \oplus a_t$ and $\mathcal{A}^{t+1}_{\mathcal{C}_\mathbf{I}} = \mathcal{A}^t_{\mathcal{C}_\mathbf{I}} \setminus a_t$, respectively. The remaining candidates are then reevaluated with respect to $\mathcal{V}_t$. The evolution terminates when no candidate improves both consistency measures or when the candidate pool is exhausted, yielding the final image-adaptive vocabulary set $\mathcal{V}^*$. We provide the pseudo-code in Sec. C.

Vocabulary Evaluation at Round t. At round t, each remaining candidate alias a is evaluated by comparing the current vocabulary set $\mathcal{V}_{t-1}$ with its candidate expansion $\mathcal{V}_{t-1} \oplus a$. We apply the same prediction procedure to both vocabulary sets. Specifically, for a vocabulary set $\mathcal{V} \in \mathcal{V}_{t-1}, \mathcal{V}_{t-1} \oplus a$, we compute alias-level[1] logits for the global and local views by matching their patch features with the corresponding text embeddings. The alias-level logits within each category are aggregated into class logits, followed by argmax and one-hot encoding to produce view-wise predictions:

$$\mathbf{Z}^{\mathcal{V}}_{v,c} = \text{Agg}\left(\{\text{cosine}\left(\mathbf{F}^H_v, \widetilde{\mathbf{T}}_{\mathbf{I},a'}\right)\}_{a' \in \mathcal{V}_c}\right), \quad \mathbf{S}^{\mathcal{V}}_v = \text{OneHot}\left(\arg\max_{c \in \mathcal{C}} [\mathbf{Z}^{\mathcal{V}}_v]_{\cdot,c}\right), \tag{7}$$

where $v \in \{G, L_1, \ldots, L_R\}$. Here, $\mathcal{V}_c$ contains the original name of category c and its selected aliases, $\mathbf{F}^H_v$ denotes the holistic visual representation of view v, and $\widetilde{\mathbf{T}}_{\mathbf{I},a'}$ denotes the text embedding of alias a' for image $\mathbf{I}$. The operator Agg aggregates the alias-level logits within each category.

The resulting paired predictions are evaluated from semantic and structural perspectives. We first measure local-global semantic consistency by comparing each local prediction with the corresponding region in the global prediction. Let $\mathbf{S}^{\mathcal{V}}_{G,r} = \mathcal{R}_r(\mathbf{S}^{\mathcal{V}}_G)$ denote the geometrically corresponding region cropped from the global prediction and resized to the spatial resolution of local view L_r. For each vocabulary set $\mathcal{V}$, its semantic consistency score is defined as:

$$\mathcal{Q}^{\mathcal{V}}_{\text{sem}} = \frac{1}{R}\sum_{r=1}^{R} \mathcal{S}\left(\mathbf{S}^{\mathcal{V}}_{L_r}, \mathbf{S}^{\mathcal{V}}_{G,r}\right), \quad \Delta_{\text{sem},t}(a) = \mathcal{Q}^{\mathcal{V}_{t-1}\oplus a}_{\text{sem}} - \mathcal{Q}^{\mathcal{V}_{t-1}}_{\text{sem}}, \tag{8}$$

where $\mathcal{S}$ measures prediction agreement using mIoU. Thus, $\Delta_{\text{sem},t}(a)$ measures the change in local-global semantic consistency after adding candidate a.

We then measure structural consistency within each local view by comparing its direct prediction with the corresponding hard prediction after random-walk refinement. Following the refinement process in Eqn. 3, the structural consistency score and the corresponding gain of candidate a are defined as:

$$\mathcal{Q}^{\mathcal{V}}_{\text{str}} = \frac{1}{R}\sum_{r=1}^{R} \mathcal{S}\left(\mathbf{S}^{\mathcal{V}}_{L_r}, \mathcal{H}\left(\left(\mathbf{A}_{L_r}\right)^{\kappa} \mathbf{S}^{\mathcal{V}}_{L_r}\right)\right), \quad \Delta_{\text{str},t}(a) = \mathcal{Q}^{\mathcal{V}_{t-1}\oplus a}_{\text{str}} - \mathcal{Q}^{\mathcal{V}_{t-1}}_{\text{str}}. \tag{9}$$

Here, $\mathcal{H}(\cdot)$ applies class-wise min-max normalization, followed by pixel-wise argmax and one-hot encoding to obtain a hard prediction map. The local prediction $\mathbf{S}^{\mathcal{V}}_{L_r}$ is obtained using Eqn. 7, while and $\mathbf{A}_{L_r}$ denotes the same transition matrix used for random-walk refinement in Eqn. 3. Therefore, $\Delta_{\text{str},t}(a)$ measures whether adding candidate a improves the agreement between the local predictions and their corresponding visual structures.

The two gains jointly evaluate candidate a. As defined in Eqn. 6, the candidate is eligible for selection only when both $\Delta_{\text{sem},t}(a)$ and $\Delta_{\text{str},t}(a)$ are positive, and the eligible candidate with the largest total gain is selected to update the vocabulary set.

[1]Categories outside $\mathcal{C}_\mathbf{I}$ remain in the vocabulary set and are represented only by their original category names.

Table 1: Comparison with state-of-the-art training-free OVSS methods across eight evaluation protocols, with results reported as mIoU (%). Avg. denotes the mean mIoU across all eight protocols. Best results are in **bold**, and second best are underlined.

Method	External Model	*With Background*			*Without Background*					Avg.
		VOC21	PC60	Object	VOC20	PC59	Stuff	City	ADE	
CLIP-based										
CLIP[ICML'21]	✗	16.2	7.7	5.5	41.8	9.2	4.4	5.5	2.1	11.6
MaskCLIP[ECCV'22]	✗	38.8	23.6	20.6	74.9	26.4	16.4	12.6	9.8	27.9
SCLIP[ECCV'24]	✗	59.1	30.4	30.5	80.4	34.2	22.4	32.2	16.1	38.2
ClearCLIP[ECCV'24]	✗	51.8	32.6	33.0	80.9	35.9	23.9	30.0	16.7	38.1
ProxyCLIP[ECCV'24]	DINO	61.3	35.3	37.5	80.3	39.1	26.5	38.1	20.2	42.3
NACLIP[WACV'25]	✗	58.9	32.2	33.2	79.7	35.2	23.3	35.5	17.4	39.4
ResCLIP[CVPR'25]	✗	60.9	33.4	34.7	85.9	36.6	24.6	35.6	18.0	41.2
FreeCP[ICCV'25]	Vicuna	64.5	35.7	36.9	81.5	39.3	26.1	34.4	18.9	42.2
Trident[ICCV'25]	DINO+SAM	67.1	38.6	41.1	84.5	42.2	28.3	42.9	21.9	45.8
CorrCLIP[ICCV'25]	DINO+SAM	74.8	44.2	43.7	88.8	48.8	31.6	49.4	26.9	51.0
PEARL[CVPR'26]	✗	64.1	35.1	37.3	86.9	38.6	26.3	37.6	19.4	43.2
DSLO[CVPR'26]	Stable Diffusion	67.8	38.3	39.6	89.0	44.9	28.9	43.3	23.0	46.9
GLA-CLIP[CVPR'26]	DINO	66.3	36.1	37.7	84.2	39.9	26.9	40.8	20.0	44.0
OV-Stitcher[CVPR'26]	DINO+SAM+LLaMA	75.7	43.9	42.6	89.8	48.8	31.8	48.1	24.7	50.7
dino.txt-based										
dino.txt[CVPR'25]	✗	35.9	22.9	24.2	84.5	26.1	19.2	22.6	18.9	31.8
VIP[ICML'26]	GPT	73.2	41.2	47.4	92.5	46.5	33.3	55.7	29.1	52.4
HIVE (Ours)	GPT	**78.7**	**47.5**	**52.9**	**94.7**	**52.0**	**34.8**	**57.5**	**31.6**	**56.2**

4 Experiments

4.1 Experimental Setup

Datasets and metrics. Following previous work (Lan et al., 2024b; Zhu et al., 2026), we evaluate HIVE on the validation splits of six datasets under eight OVSS evaluation settings, using mean intersection over union (mIoU, %). These datasets include PASCAL VOC (VOC20/VOC21) (Everingham et al., 2015), PASCAL Context (PC59/PC60) (Mottaghi et al., 2014), COCO-Object (Object) (Caesar et al., 2018), COCO-Stuff164k (Stuff) (Caesar et al., 2018), Cityscapes (City) (Cordts et al., 2016), and ADE20K (ADE) (Zhou et al., 2019). VOC21, PC60, and Object include a background category, while the remaining settings exclude it.

Implementation details. Our HIVE is built upon DINOv3-based dino.txt model (Jose et al., 2025; Siméoni et al., 2025). All pretrained parameters in HIVE remain frozen during inference. We use GPT-5.6 Sol (OpenAI, 2026) to generate alias candidates with the prompt, demonstrated in Sec. B for detail. Results with other LLMs are reported in Sec. F. Following prior work (Shi et al., 2025), we resize the image shorter side to 336 pixels with keeping ratio aspect for all datasets except Cityscapes, where it is set to 560 pixels. We extract overlapping 336^2 local views with a stride of 112 pixels for all datasets. The global view is resized to 336^2, and all views are independently processed by the visual encoder $\mathscr{E}$. We set B=3, $\mathcal{K}$=2, which are evaluated in Sec. D E. All experiments are conducted on a workstation equipped with an NVIDIA RTX A40 GPU. More details are in Sec. A.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares HIVE with 16 training-free OVSS methods across eight evaluation protocols. HIVE achieves an average mIoU of 56.2%, outperforming dino.txt (Jose et al., 2025) by 24.4%. HIVE also achieves higher average mIoU than SCLIP (Wang et al., 2024) and ClearCLIP (Lan et al., 2024a), which refine visual representations through attention reconstruction and residual removal, respectively. Compared with methods that incorporate external spatial priors, including Prox-

Table 2: Components evaluation, with results reported by mIoU (%). Avg. denotes the mean mIoU across four datasets protocol.

ID	Method	RandomWalk	HVR	IVE	VOC21	Object	City	PC59	Avg.
(a)	Baseline	×	×	×	67.2	47.0	41.0	41.7	49.2
(b)	RandomWalk	✓	×	×	70.7	48.8	42.1	43.9	51.4
(c)	HVR	×	✓	×	69.5	48.0	42.6	43.2	50.8
(d)	IVE	×	×	✓	73.4	50.1	51.7	47.7	55.7
(e)	HIVE (ours)	✓	✓	✓	**78.7**	**52.9**	**57.5**	**52.0**	**60.3**

Table 3: Evaluation of HVR in terms of mIoU (%). Avg. denotes the mean mIoU across the four evaluation protocols.

ID	Local-view	Global-view	Refinement		VOC21	Object	City	PC59	Avg.
			Cross-view	Scheme					
(a)	✓	×	×	–	67.2	47.0	41.0	41.7	49.2
(f)	✓	✓	×	Mean	64.1	44.5	36.2	39.4	46.1
(g)	✓	×	✓	L2L	68.3	47.2	41.5	42.5	49.9
(h)	✓	✓	✓	G2L	68.9	47.6	42.0	42.7	50.3
(c)	✓	✓	✓	L2G2L	**69.5**	**48.0**	**42.6**	**43.2**	**50.8**

Table 4: Evaluation of IVE in terms of mIoU (%). Avg. denotes the mean mIoU across the four evaluation protocols. $\mathcal{C}$, $\mathcal{C}_E$, $\mathcal{V}_0$, and $\mathcal{V}^*$ denote the original category names, image-shared expanded vocabulary, image-adaptive expanded vocabulary before evolution, and evolved image-adaptive vocabulary, respectively.

ID	Vocabulary	Expanded	Image-adaptive	Evolution	VOC21	Object	City	PC59	Avg.
(a)	$\mathcal{C}$	×	×	–	67.2	47.0	41.0	41.7	49.2
(i)	$\mathcal{C}_E$	✓	×	–	65.7	45.6	43.9	40.5	48.9
(j)	$\mathcal{V}_0$	✓	✓	×	68.5	47.7	45.3	42.8	51.1
(k)	$\mathcal{V}^*$	✓	✓	Random	67.2	47.3	44.8	41.6	50.2
(l)	$\mathcal{V}^*$	✓	✓	One-shot	71.9	48.4	48.7	45.6	53.7
(d)	$\mathcal{V}^*$	✓	✓	Iterative	**73.4**	**50.1**	**51.7**	**47.7**	**55.7**

yCLIP (Lan et al., 2024b) and Trident (Shi et al., 2025), HIVE maintains an advantage, exceeding Trident by 10.4% in average mIoU. It further surpasses the cross-window methods GLA-CLIP (Lee et al., 2026) and OV-Stitcher (Moon et al., 2026) by 12.2% and 5.5%, respectively. Beyond methods focused on visual representations, HIVE improves over the text prompt refinement methods FreeCP (Chen et al., 2025) and VIP (Zhu et al., 2026) by 14.0% and 3.8% in average mIoU. The gains over VIP are mainly consistent with the benefits of adapting the vocabulary to each image rather than using a vocabulary shared across images.

4.3 Ablation Studies

In the following, we provide an in-depth discussion of our HIVE. Specifically, we first evaluate the individual and joint contributions of the components in Table 2, and then evaluate the designs in HVR and IVE in Tables 3 and 4, respectively. Results are reported on four evaluation protocols (VOC21, Object, City, and PC59), and we also report the average mIoU (Avg.) across them. The best performance is shown in **bold**.

Component Evaluation. Table 2 evaluates the contributions of HVR, IVE and random-walk refinement. Random-walk refinement improves performance across all four evaluation protocols, increasing average mIoU from 49.2% to 51.4%. Furthermore, HVR or IVE individually provides further gains of 1.6% and 6.5%, respectively. Combining HVR and IVE raises average mIoU to 60.3%, exceeding the configurations with HVR alone and IVE alone by 9.5% and 4.6%, respectively. These additional gains support the joint use of the two modules to improve dense prediction.

Evaluation of HVR. Table 3 evaluates different view configurations and refinement schemes for HVR. The first panel focuses on intra-view refinement, showing that directly incorporating the coarse global-view predictions into prediction fusion in (f) leads to a performance drop of 3.1%.

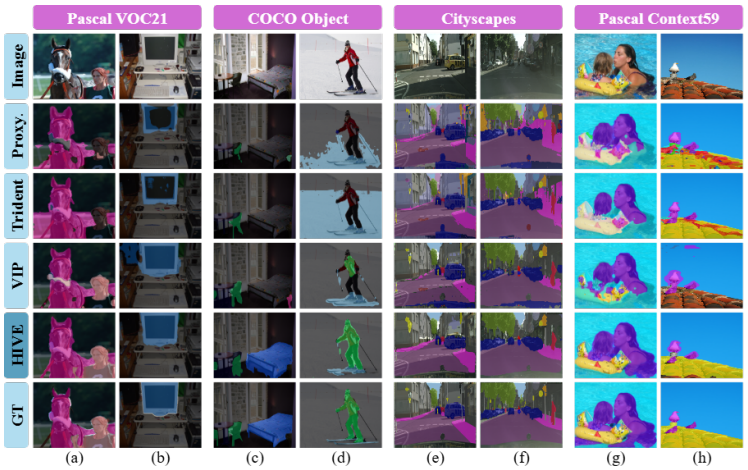


Figure 4: Qualitative comparison with counterparts on four datasets protocol.

In the second panel, we evaluate three cross-view refinement schemes. Compared with the intra-view baseline, direct local-to-local interaction (L2L) in (g), global-to-local interaction (G2L) in (h), and the proposed local-to-global-to-local interaction (L2G2L) in (c) yield performance gains of 0.7%, 1.1%, and 1.6%, respectively. Furthermore, these results demonstrate the effectiveness of incorporating local cues to refine the global representation before cross-view interaction.

Evaluation of IVE. Table 4 evaluates semantic expansion, image-adaptive vocabulary construction, and different vocabulary evolution strategies. Without vocabulary evolution setting presented in first panel, category-level aliases shared across images $\mathcal{C}_E$ in (i) performs better than the original category names $\mathcal{C}$ in (a), but inferior to the image-adaptive expanded vocabulary $\mathcal{V}_0$ in (j). We further evaluate different evolution strategies by given $\mathcal{V}_0$ in the second panel. Randomly selecting a fixed number of aliases in (k) yields performance comparable to directly using $\mathcal{V}_0$ in (k). In (l), one-shot selection computes the gain of each candidate alias with respect to $\mathcal{V}_0$ and adds the top-ranked aliases in a single update. It outperforms random selection by 3.5% but falls behind our iterative policy in (d) by 2%. This result supports the iterative reevaluation of candidates after each update.

4.4 Qualitative Analysis

Figure 4 compares HIVE with some representative training-free OVSS methods across multiple datasets. HIVE produces more complete masks with clearer boundaries and more accurate semantic labels. The contribution of HVR is evident in fine-grained regions. For example, HIVE better preserves the small object in Fig. 4(h) and recovers the complete boundary of object in Fig. 4(b), whereas other methods produce coarse predictions. IVE mainly improves semantic recognition by providing image-specific textual evidence. In Fig. 4(c), competing methods confuse object with incorrect category, while HIVE correctly identifies it as correct category. These comparisons demonstrate that HVR improves fine-grained grounding, while IVE strengthens semantic alignment.

5 Conclusion

In this work, we introduce HIVE to jointly improve visual representations and textual vocabularies for training-free open-vocabulary semantic segmentation. To obtain fine-grained visual descriptors for dense prediction, the Holistic Visual Reasoning (HVR) module enhances local features by modeling cross-view relations. It first refines the coarse global representation with local cues and then uses the refined representation as a bridge across local views. The Image-adaptive Vocabulary Evolution (IVE) module uses an MLLM to construct an image-specific pool of category aliases and iteratively evolves the vocabulary according to image semantics and visual structure, yielding a dedicated vocabulary for each image. Experiments on eight evaluation protocols show that HIVE achieves state-of-the-art performance. The results validate the effectiveness of local feature refinement and image-adaptive vocabulary evolution for training-free OVSS.

AI USE STATEMENT

The research ideas and methodology presented in this work were independently conceived and designed by the authors. As part of the proposed Image-adaptive Vocabulary Evolution (IVE) module, an off-the-shelf LLM (GPT) is employed solely for text augmentation. For manuscript preparation, generative AI tools were used exclusively for grammar checking and improving language clarity. The authors reviewed all AI-assisted outputs and take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work uses publicly available datasets and pretrained models and does not collect new human-subject or personally identifiable data. The MLLM used for alias generation may inherit biases or produce inaccurate descriptions, which could affect segmentation performance across categories or visual contexts. We mitigate this risk by enforcing class-preserving alias generation, retaining all original category names, and selecting aliases according to visual consistency. As with other semantic segmentation techniques, the proposed method could potentially be misused in privacy-sensitive applications such as surveillance. We encourage its responsible use in compliance with applicable dataset licenses, privacy requirements, and ethical standards.

REPRODUCIBILITY STATEMENT

All experiments in this work use publicly available datasets and pretrained models (Jose et al., 2025). We evaluate HIVE on eight protocols derived from six datasets: PASCAL VOC (VOC20/VOC21) (Everingham et al., 2015), PASCAL Context (PC59/PC60) (Mottaghi et al., 2014), COCO-Object (Object) (Caesar et al., 2018), COCO-Stuff164k (Stuff) (Caesar et al., 2018), Cityscapes (City) (Cordts et al., 2016), and ADE20K (ADE) (Zhou et al., 2019). The main paper provides the complete formulation and architecture of HIVE, while the supplementary material includes implementation details, hyperparameter settings, the MLLM prompt used for alias generation, and the pseudocode for vocabulary evolution. Upon acceptance, we will publicly release the source code and the generated image-specific vocabulary sets to facilitate reproduction and further research.

REFERENCES

Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218. IEEE, 2018.

Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 23124–23134, 2025.

Zhixiang Chi, Yanan Wu, Li Gu, Huan Liu, Ziqiang Wang, Yang Zhang, Yang Wang, and Konstantinos Plataniotis. Plug-in feedback self-adaptive attention in CLIP for training-free open-vocabulary segmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 22815–22825, 2025.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, and Paul Hongsuck Seo. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4123, 2024.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pp. 540–557, 2022.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5061–5071, 2025.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. DINOv2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24905–24916, 2025.

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 *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *European Conference on Computer Vision*, pp. 143–160, 2024a.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *European Conference on Computer Vision*, pp. 70–88, 2024b.

ByeongCheol Lee, Hyun Seok Seong, Sangeek Hyun, Gilhan Park, WonJun Moon, and Jae-Pil Heo. Looking beyond the window: Global-local aligned CLIP for training-free open-vocabulary semantic segmentation. *arXiv preprint arXiv:2603.23030*, 2026.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations*, 2023.

Seungjae Moon, Seunghyun Oh, and Youngmin Ro. OV-Stitcher: A global context-aware framework for training-free open-vocabulary semantic segmentation. *arXiv preprint arXiv:2604.08110*, 2026.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 891–898, 2014.

OpenAI. GPT-5.6 System Card. OpenAI Deployment Safety Hub, July 2026. Published July 9, 2026.

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 *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23487–23497. IEEE, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Wenfang Sun, Yingjun Du, Gaowen Liu, Ramana Kompella, and Cees G. M. Snoek. Training-free semantic segmentation via LLM-supervision. *arXiv preprint arXiv:2404.00701*, 2024.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pp. 315–332, 2024.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. ResCLIP: Residual attention for training-free dense vision-language inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29968–29978, 2025.

Bingfeng Zhang, Siyue Yu, Hui Li, Jiahua Lin, Wenwu Wang, and Jimin Xiao. The power of prior: Training-free open-vocabulary semantic segmentation with LLaVA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International journal of computer vision*, 127(3):302–321, 2019.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision*, pp. 696–712, 2022.

Chufeng Zhou, Jian Wang, Xinyuan Liu, Jie Zheng, and Xiaokang Zhang. Geospatial-reasoning-driven vocabulary-agnostic remote sensing semantic segmentation. *arXiv preprint arXiv:2602.08206*, 2026.

Hao Zhu, Shuo Jin, Wenbin Liao, Jiayu Xiao, Yan Zhu, Siyue Yu, and Feng Dai. VIP: Visual-guided prompt evolution for efficient dense vision-language inference. In *Forty-third International Conference on Machine Learning*, 2026.

APPENDIX

This appendix provides additional implementation details, methodology demonstration, and experimental analyses for HIVE. The contents are organized as follows:

- Experimental details in Sec. A;
- Prompt design in Sec. B;
- Pseudocode for Image-adaptive Vocabulary Evolution in Sec. C;
- Effect of the number of generated aliases in Sec. D;
- Effect of the number of random-walk steps in Sec. E;
- Comparison of different LLMs in Sec. F;
- Limitations and future work in Sec. G.

A EXPERIMENTAL DETAILS

Our baseline is built upon the DINOv3-based dino.txt model (Jose et al., 2025; Siméoni et al., 2025). It comprises a DINOv3 ViT-L/16 visual encoder $\mathscr{E}$ followed by a text-alignment adapter $\mathscr{A}$, together with a pretrained text encoder. Given a set of local view images as input, it summarizes their predictions as the final output. In HIVE, we modify the attention layers of $\mathscr{A}$ to implement feature-wise HVR. We set $\beta = 1.5$ and $\gamma = 0.5$ in practice.

B IMPLEMENTATION OF PROMPTS

```
Inspect the image and this numbered initial vocabulary:[{"id":0,"people":"..."},{"id":1,"sofa":"..."},
...]>

The vocabulary terms are mutually exclusive categories. The same visible entity or region must not
be assigned to two different IDs. If two terms could describe the same visual evidence, choose
only the single best-supported ID, or omit it when ambiguous.
Do not count a component of one object as a separate category merely because that component
resembles another vocabulary term. For example, a car's doors and windows are parts of the car,
not evidence for the door or window categories. Select a door or window only when it is visibly a
distinct scene element, such as an architectural door or window. Apply this part-versus-category
distinction to all terms.
Return only IDs for categories you are highly confident are visibly present. Omit uncertain, barely
visible, inferred, or ambiguous categories.
For each selected ID, write five short English aliases that accurately name that category as it
appears in this image. Prefer familiar synonyms, visible subtypes, plurals, or simple visually
supported modifiers. Do not name related objects or describe relationships. Do not use an
object's components as its aliases, or write an alias that names a different vocabulary category.
Never invent appearance details.
No alias may equal any original vocabulary term (case-insensitive). No alias may appear under two
IDs (case-insensitive).
Return only JSON with a categories array of objects containing exactly an integer id and an
aliases array. Omit unselected IDs. Do not return category names, explanations, or other fields.
```

Figure 5: Prompt used to query GPT for visible-category identification and image-conditioned alias generation.

Figure 5 presents the complete prompt used for image-adaptive semantic expansion. Given an input image and a numbered initial vocabulary, GPT first identifies the categories that are confidently visible. To reduce semantic ambiguity, the prompt enforces mutually exclusive category assignments, distinguishes standalone scene elements from object components, and excludes uncertain or inferred categories. For each selected category, GPT generates three ($B = 3$) short English aliases restricted to class-preserving synonyms, visible subtypes, plurals, or visually supported modifiers. Original category names, aliases shared across different category IDs, unsupported appearance descriptions, and terms belonging to other categories are prohibited. The response is constrained to a JSON object

Input: $\mathbf{I}, \mathcal{C}, \mathbf{F}_G^H, \{\mathbf{F}_{L_r}^H, \mathbf{A}_{L_r}\}_{r=1}^R, B$.
Output: $\mathcal{V}^*$.

Generate $\mathcal{A}_{\mathcal{C}_\mathbf{I}}$ using the MLLM — *# Semantic expansion*
$\mathcal{V}_0 \leftarrow \mathcal{C}, \quad t \leftarrow 1$ — *# Initial vocabulary*
$\mathcal{A}_{\mathcal{C}_\mathbf{I}}^1 \leftarrow \mathcal{A}_{\mathcal{C}_\mathbf{I}} \setminus \mathcal{V}_0$ — *# Remaining aliases*
Obtain $\mathbf{S}_G^{\mathcal{V}_0}$ and $\{\mathbf{S}_{L_r}^{\mathcal{V}_0}\}_{r=1}^R$ using Eqn. equation 7
Compute $\mathcal{Q}_{\text{sem}}^{\mathcal{V}_0}$ using Eqn. equation 8
Obtain $\{\widetilde{\mathbf{S}}_{L_r}^{\mathcal{V}_0}\}_{r=1}^R$ using Eqn. equation 3
Compute $\mathcal{Q}_{\text{str}}^{\mathcal{V}_0}$ using Eqn. equation 9
while $\mathcal{A}_{\mathcal{C}_\mathbf{I}}^t \neq \varnothing$ **do** — *# For round t*
  foreach $a \in \mathcal{A}_{\mathcal{C}_\mathbf{I}}^t$ **do**
    Obtain $\mathbf{S}_G^{\mathcal{V}_{t-1}\oplus a}$ and $\{\mathbf{S}_{L_r}^{\mathcal{V}_{t-1}\oplus a}\}_{r=1}^R$ using Eqn. equation 7
    Compute $\mathcal{Q}_{\text{sem}}^{\mathcal{V}_{t-1}\oplus a}$ using Eqn. equation 8
    Obtain $\{\widetilde{\mathbf{S}}_{L_r}^{\mathcal{V}_{t-1}\oplus a}\}_{r=1}^R$ using Eqn. equation 3
    Compute $\mathcal{Q}_{\text{str}}^{\mathcal{V}_{t-1}\oplus a}$ using Eqn. equation 9
    $\Delta_{\text{sem},t}(a) \leftarrow \mathcal{Q}_{\text{sem}}^{\mathcal{V}_{t-1}\oplus a} - \mathcal{Q}_{\text{sem}}^{\mathcal{V}_{t-1}}$
    $\Delta_{\text{str},t}(a) \leftarrow \mathcal{Q}_{\text{str}}^{\mathcal{V}_{t-1}\oplus a} - \mathcal{Q}_{\text{str}}^{\mathcal{V}_{t-1}}$
    $\Delta_t(a) \leftarrow \Delta_{\text{sem},t}(a) + \Delta_{\text{str},t}(a)$ — *# Total gain*
  if $\nexists\, a \in \mathcal{A}_{\mathcal{C}_\mathbf{I}}^t$ *such that* $\Delta_{\text{sem},t}(a) > 0$ *and* $\Delta_{\text{str},t}(a) > 0$ **then**
    break — *# No eligible alias*
  $a_t \leftarrow \underset{\substack{a\in\mathcal{A}_{\mathcal{C}_\mathbf{I}}^t \\ \Delta_{\text{sem},t}(a)>0 \\ \Delta_{\text{str},t}(a)>0}}{\arg\max}\ \Delta_t(a)$ — *# Eqn. equation* **??**
  $\mathcal{V}_t \leftarrow \mathcal{V}_{t-1} \oplus a_t$ — *# Update vocabulary*
  $\mathcal{A}_{\mathcal{C}_\mathbf{I}}^{t+1} \leftarrow \mathcal{A}_{\mathcal{C}_\mathbf{I}}^t \setminus \{a_t\}$ — *# Remove selected alias*
  $\mathcal{Q}_{\text{sem}}^{\mathcal{V}_t} \leftarrow \mathcal{Q}_{\text{sem}}^{\mathcal{V}_{t-1}\oplus a_t}$
  $\mathcal{Q}_{\text{str}}^{\mathcal{V}_t} \leftarrow \mathcal{Q}_{\text{str}}^{\mathcal{V}_{t-1}\oplus a_t}$
  $t \leftarrow t + 1$
$\mathcal{V}^* \leftarrow \mathcal{V}_{t-1}$
return $\mathcal{V}^*$

Algorithm 1: Pseudo-code for Image-adaptive Vocabulary Evolution.

containing only the selected category IDs and their aliases, allowing the candidate pool $\mathcal{A}_{\mathcal{C}_\mathbf{I}}$ to be parsed directly.

C Pseudocode for Image-adaptive Vocabulary Evolution

Algorithm 1 summarizes the vocabulary evolution procedure, which is performed independently for each image. We initialize the vocabulary set as $\mathcal{V}_0 = \mathcal{C}$ and progressively add aliases from the MLLM-generated candidate pool. All original category names are retained, while the visual representations and local transition matrices remain fixed throughout the search.

At round t, we first compute the global and local predictions of the current vocabulary set $\mathcal{V}_{t-1}$ and evaluate its semantic and structural consistency scores. For each remaining alias a, we apply the same procedure to the expanded vocabulary set $\mathcal{V}_{t-1} \oplus a$. The semantic and structural gains are obtained by comparing the corresponding scores before and after adding a. An alias is eligible only when both gains are positive, and the eligible alias with the largest total gain is added to the vocabulary set. The remaining aliases are then reevaluated with respect to the updated vocabulary set. The procedure terminates when no eligible alias remains or the candidate pool is exhausted.

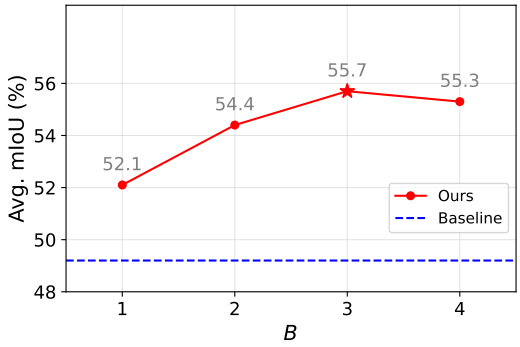


Figure 6: Evaluation on generated aliases B per visible category. The dashed line indicates the baseline without IVE. Avg.: The mean mIoU across VOC21, Object, City and PC59 datasets.

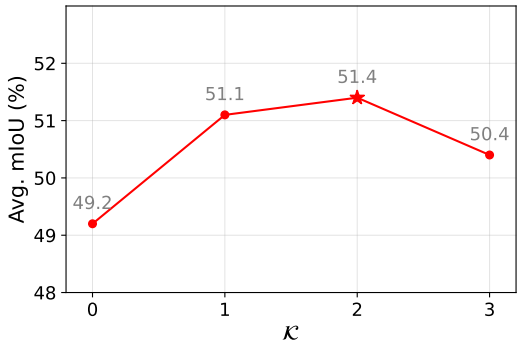


Figure 7: Evaluation on random-walk steps $\mathcal{K}$. Avg.: The mean mIoU across VOC21, Object, City and PC59 datasets.

D Evaluation of Vocabulary Number B

Figure 6 evaluates the number of aliases B generated for each visible category under the baseline setting. Performance improves steadily as B increases, indicating that a more diverse candidate pool benefits image-adaptive vocabulary evolution. The best average mIoU of 55.7% is achieved at $B = 3$. Further increasing B leads to a slight performance decline, suggesting that excessive candidates may introduce redundant or less relevant semantics. All evaluated settings outperform the baseline, and we therefore set $B = 3$ in all experiments.

E Evaluation of Random Walk Step Number $\mathcal{K}$

Figure 7 evaluates the number of random-walk steps $\mathcal{K}$ using the average mIoU across the four ablation protocols. Here, $\mathcal{K} = 0$ denotes the baseline without random walk. Introducing one propagation step improves the average mIoU from 49.2% to 51.1%, demonstrating the benefit of exploiting local visual structure for prediction refinement. Performance further increases to 51.4% at $\mathcal{K} = 2$, yielding a 1.9% improvement over the baseline. Increasing $\mathcal{K}$ to 3 reduces the average mIoU to 50.4%. We therefore set $\mathcal{K} = 2$ in all experiments.

F Comparison of Various LLMs

Table 5 evaluates the influence of different LLMs on IVE. All HIVE variants using different LLMs substantially outperform the baseline, demonstrating the general effectiveness of image-adaptive semantic expansion. Among them, GPT achieves the best average mIoU of 55.7%, exceeding the baseline by 6.5% and outperforming Qwen-3.7 and DeepSeek-V4 by 0.8% and 1.1%, respectively.

Table 5: Evaluation of various LLM in terms of mIoU (%). Avg. denotes the mean mIoU across the four evaluation protocols.

Method	LLM	VOC21	Object	City	PC59	Avg.
Baseline	—	67.2	47.0	41.0	41.7	49.2
HIVE (ours)	Qwen-3.7	72.8	48.7	50.8	47.1	54.9
HIVE (ours)	DeepSeek-V4	72.4	48.2	51.3	46.4	54.6
HIVE (ours)	GPT-5.6 Sol	**73.4**	**50.1**	**51.7**	**47.7**	**55.7**

G Limitation and Future Work

In this work, IVE requires iterative evaluation of candidate aliases for each image, introducing additional inference cost as the candidate pool grows. This cost could be reduced through parallel candidate evaluation and cached text embeddings. The resulting vocabulary also depends on the visible categories and aliases generated by the MLLM. Candidate verification or uncertainty-aware generation could improve robustness to missing categories and noisy aliases. Future work could extend HIVE to semi-supervised or weakly supervised segmentation to exploit limited annotations. Additionally, the local-to-global-to-local interaction could be extended from feature representations to affinity matrices, enabling the model to capture structural relations across views and provide more reliable guidance for prediction refinement.