REGDEC: DECOUPLING REGION CUES FROM GLOBAL SEMANTICS FOR VLM-GROUNDED OPEN-VOCABULARY DENSE PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Pretrained Models (VLMs) achieve strong open-vocabulary perception by aligning visual representations with natural language semantics, but their global image-text pre-training objectives often limit local discriminability for dense perception. Existing approaches improve dense perception through feature refinement, external spatial guidance, or task-specific adaptation. In this paper, we analyze how spatial visual cues evolve across VLM depths and observe that preferred feature depths vary across tasks and regions. These observations motivate region-adaptive fusion of semantic and region cues across layers. To address this, we propose **ReGDec** (**Re**gion-**G**lobal **Dec**oupling), a region-adaptive framework that decomposes region features into CLS-aligned components and orthogonal residuals. A compact selector learns separate region-conditioned weights to combine the two components across layers, while keeping the VLM encoders frozen. Experiments across multiple VLM backbones and dense perception benchmarks demonstrate the effectiveness of ReGDec, with gains varying across backbones and tasks. The anonymous code repository is available at: `https://anonymous.4open.science/r/ReGDec-6F5C/`.

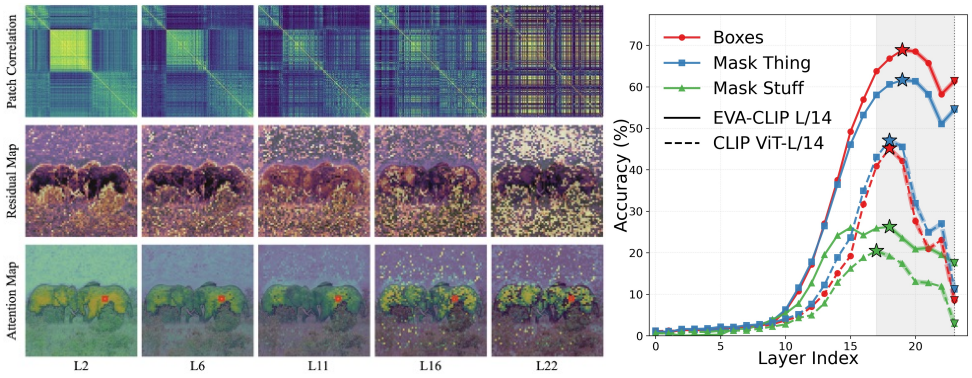
1 INTRODUCTION

Vision-Language Pretrained Models (VLMs), such as the CLIP family (Radford et al., 2021; Cherti et al., 2023; Chuang et al., 2025), have achieved remarkable generalization across diverse downstream tasks, including zero-shot image classification (Jia et al., 2021; Radford et al., 2021), visual question answering (Antol et al., 2015; Khan et al., 2022; Yu et al., 2022), and image-text retrieval (Cho et al., 2021; Li et al., 2022b; Mishra et al., 2013), by aligning visual representations with natural language semantics. Their success has also stimulated the emergence of open-vocabulary dense perception tasks (Wang et al., 2025; Qiu et al., 2025), such as Open-Vocabulary Object Detection (OVOD) and Open-Vocabulary Semantic Segmentation (OVSS), which require VLMs to perform both global image-text understanding and region perception. This poses a challenge for VLMs, whose pre-training objectives primarily optimize global image-text alignment, leaving their region representations insufficiently discriminative for dense perception tasks.

To this end, two common approaches improve region representations through external guidance or dense adaptation. The first (Wang et al., 2025; Zhang et al., 2025) injects external model guidance, e.g., Vision Foundation Models (VFMs) (Caron et al., 2021; Oquab et al., 2023; Ravi et al., 2024; Carion et al., 2025), whose strong spatial perception complements the global image-text alignment of VLMs. The second leverages additional supervision, either via dense-specific finetuning (Wu et al., 2023b; Qiu et al., 2025) or knowledge distillation from auxiliary teachers (Wysoczańska et al., 2024), to obtain discriminative dense representations. Another line of work recovers local information directly from frozen CLIP representations (Zhou et al., 2022a; Li et al., 2025; Bousselham et al., 2024; Hajimiri et al., 2025). Building on these findings, we investigate how the utility of region features varies across layers and regions, and how semantic and region components from different depths can be adaptively combined.

To investigate this, we trace the layer-wise evolution of attention patterns and patch correlations. As shown in Fig. 1(a), patch correlations and attention maps in intermediate layers preserve object

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.



(a) Layer-wise object attention. (b) Layer-wise accuracy.

Figure 1: Region cues are entangled with global semantics and task-dependent in VLM depth. (a) Object structure is preserved at mid layers but diffuses to global context at later layers; (b) per-task accuracy peaks (stars) before the final layer at task-dependent depths.

structure, but diffuse to broader global contexts at later layers. As a result, region cues are partially weakened during late-stage global semantic aggregation, leading to reduced region-level discriminability of patch representations. This observation indicates that region cues should be decoupled from global text-aligned semantics and adaptively recombined to enhance region discriminability without compromising VLMs' semantic capability. Motivated by this, we conduct dense perception experiments with controlled region cue selection in Fig. 1(b), finding that properly choosing global semantics with region cues substantially improve accuracy.

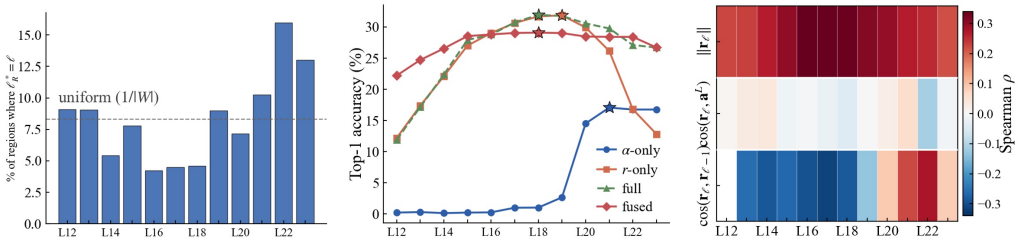
However, further observations show that the preferred feature depth varies across tasks and regions. As shown in Fig. 1(b), different tasks peak at different layers before the final one, and the per-region analysis in the following Motivation section further shows that the optimal depth varies region by region within each task. These observations motivate region-adaptive fusion of semantic and region cues across layers. Therefore, the key challenge is: *How to decouple region cues from global semantics and recombine them across VLM depths in a region-adaptive manner, while preserving the text-aligned semantic capability of VLMs?*

To address this challenge, we propose **ReGDec** (**Re**gion-**G**lobal **Dec**oupling), a region-adaptive plug-in module that decouples region cues from global semantics and recombines them for dense perception. ReGDec uses a lightweight region-conditioned *selector* to adaptively select semantic components and region cues from appropriate feature depths, enhancing region-level discriminability while preserving the text-aligned semantic capability of VLMs. ReGDec is trained on regions by updating only the selector, with the VLM encoders kept frozen. The module is then plugged into downstream pipelines without retraining their task-specific parameters. Experiments across VLM backbones and dense perception benchmarks show that ReGDec improves dense representations and achieves +9.2 AP_{50}^{novel} on OV-COCO over the baseline. Our contributions are as follows.

- We analyze how region cues vary across depths and how preferred feature depths differ across tasks and regions, motivating region-adaptive cross-layer fusion for dense perception.
- We propose **ReGDec**, which decomposes region features into CLS-aligned components and orthogonal residuals, and learns separate region-conditioned weights to combine the two components across layers.
- Experiments across CLIP-family backbones and dense recognition, detection, and segmentation benchmarks evaluate the transferability of ReGDec and characterize its gains across different backbone and readout settings.

2 Motivation

VLM representations retain region cues, yet these cues are gradually weakened by CLIP-style global image-text alignment. Without explicit region-text supervision, patch tokens are progressively shaped



(a) Preferred-layer distribution. (b) Semantic–region fusion effect. (c) Residual correlates with utility.

Figure 2: Region cues are layer-dependent and admit label-free utility prediction. (a) Region-specific optimal layers ℓ_R^* are broadly distributed across W. (b) Comparison of layer-wise CLS-aligned components a_R^ℓ, residuals r_R^ℓ, their same-layer sum, and fusion with the fixed final-layer anchor $a_R^L + r_R^\ell$. Fusion benefits vary across layers. (c) Class-agnostic statistics of r_R^ℓ correlate with the oracle utility $u_R(\ell)$ across candidate layers.

toward global semantics, suppressing patch-specific distinctions. To characterize the region-level information, we analyze patch representations geometrically by decomposing patch token into a [CLS]-aligned component and an orthogonal residual. Since this residual captures patch-specific variation useful for distinguishing regions, we refer to its region-pooled form as a *region cue*.

Patch-Level Orthogonal Decomposition. At layer ℓ, let z_i^ℓ denote the i-th patch token, c^ℓ the [CLS] token, and $\hat{c}^\ell = c^\ell / \|c^\ell\|$ its unit direction. We decompose z_i^ℓ as below:

$$z_i^\ell = \alpha_i^\ell \hat{c}^\ell + r_i^\ell \;=\; a_i^\ell + r_i^\ell, \qquad \alpha_i^\ell = z_i^{\ell\top} \hat{c}^\ell, \qquad a_i^\ell = \alpha_i^\ell \hat{c}^\ell, \qquad r_i^\ell \perp \hat{c}^\ell. \tag{1}$$

where α_i^ℓ is the scalar [CLS]-aligned projection, a_i^ℓ is the [CLS]-aligned component and r_i^ℓ is the orthogonal residual that retains patch-specific variation.

Layer-Specific Region Utility. We lift the decomposition to a region R by pooling its patches into a region [CLS]-aligned component a_R^ℓ and a region cue r_R^ℓ.

$$a_R^\ell = \tfrac{1}{|R|} \sum\nolimits_{i \in R} a_i^\ell, \quad r_R^\ell = \tfrac{1}{|R|} \sum\nolimits_{i \in R} r_i^\ell. \tag{2}$$

Let W be a candidate set of feature layers and a_R^L the text-aligned semantic reference at the final layer L. For each $\ell \in W$, we form a single-layer probe $z_{R,\ell} = a_R^L + r_R^\ell$, and score it against the frozen text embeddings $T \in \mathbb{R}^{C \times D}$ to obtain logits $\text{logits}_{R,\ell} = \text{Normalize}(z_{R,\ell})\, T^\top$. We then define the *oracle utility* of layer ℓ as the classification margin between the ground-truth and the strongest non-ground-truth category:

$$u_R(\ell) = \text{logits}_{R,\ell}[y_R] - \max_{c \neq y_R} \text{logits}_{R,\ell}[c], \qquad \ell_R^* = \arg\max_{\ell \in W} u_R(\ell). \tag{3}$$

The region-specific optimal layer ℓ_R^* identifies which candidate layer's region cue yields the largest classification margin for region R. This utility can be used as a diagnostic, with three properties:

(i) ℓ_R^* is broadly distributed across W (Fig. 2(a)), indicating that no fixed layer is universally optimal.

(ii) Varying r_R^ℓ while holding a_R^L fixed yields non-trivial accuracy across layers, whereas the [CLS]-aligned component alone is much less effective (Fig. 2(b)).

(iii) Class-agnostic statistics of r_R^ℓ, e.g., norm $\|r_R^\ell\|$, similarity to the anchor $\cos(r_R^\ell, a_R^L)$, and cross-layer coherence $\cos(r_R^\ell, r_R^{\ell-1})$, correlate consistently with $u_R(\ell)$ (Fig. 2(c)), providing label-free signals for estimating layer utility at inference.

Insufficient Single-Layer Fusion. We compare two simplified strategies. First, fusing only the residual at the anchor layer ($a_R^L + r_R^L$) yields

Table 1: Single-layer fusion, Top-1 (%).

Setup	Boxes	Mask Thing	Mask Stuff
baseline	56.7	57.9	21.0
$a_R^L + r_R^L$	58.8	60.2	18.9
$a_R^L + r_R^{\ell^*}$	**65.7**	**68.7**	<u>20.0</u>

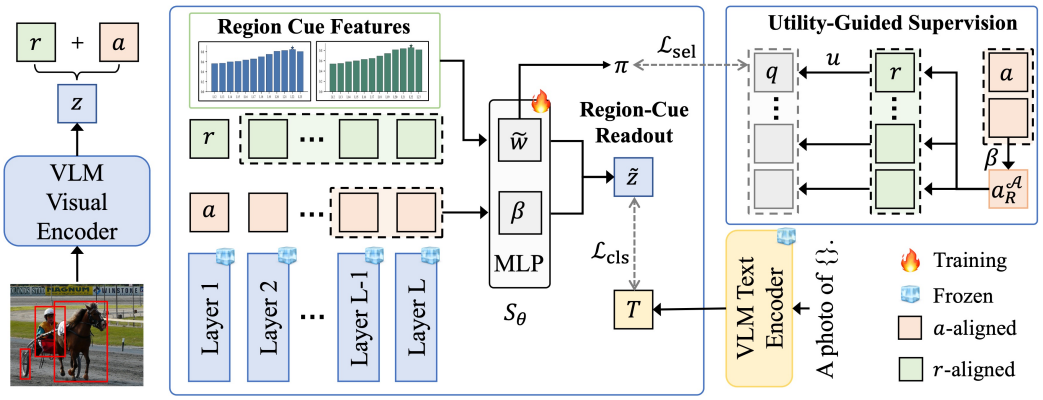


Figure 3: Region representations from the frozen VLM are decomposed into global semantic components a and orthogonal residuals r. A lightweight selector S_θ predicts semantic fusion weights and region-cue preferences from region features and geometric statistics. ReGDec then recombines selected semantic components and sparse region cues to construct the final dense representation.

modest gains on thing classes but *drops below the baseline on stuff* (18.9 vs. 21.0). Second, even the best single layer ℓ^* chosen per metric still fails to improve over the baseline on stuff ($20.0 < 21.0$, Tab. 1). This motivates a region-adaptive multi-layer fusion rather than any fixed-layer choice. Appendix A.1 provides a detailed comparison of fixed single-layer fusion and adaptive multi-layer composition.

3 Method

We instantiate **ReGDec** as a lightweight region-conditioned decoupling module on top of a frozen VLM (Fig. 3). For each region R, ReGDec separates its representation into a global text-aligned semantic component and an orthogonal residual component. It then adaptively recombines these two components: the semantic component preserves the text-aligned capability of the VLM, while the orthogonal cues enhance region-level discriminability.

3.1 Region-Cue Readout

For region R, ReGDec recombines the decomposed semantic component a_R^ℓ and residual r_R^ℓ at multiple layers. Let $\mathcal{A}$ and W be the candidate layer sets for a_R^ℓ and r_R^ℓ. The semantic reference a_R^L used in Sec. 2 corresponds to the special case $\mathcal{A} = \{L\}$. ReGDec produces the recombined region representation:

$$\tilde{z}_R \;=\; \sum\nolimits_{\ell\in\mathcal{A}} \beta_R^\ell\, a_R^\ell \;+\; \sum\nolimits_{\ell\in\mathcal{K}_R} \tilde{w}_R^\ell\, r_R^\ell, \qquad \mathcal{K}_R = \mathrm{TopK}(\pi_R, K). \tag{4}$$

Here β_R^ℓ are the softmax-normalized weights over $\mathcal{A}$ that aggregate the anchor.

Region Candidate Rescaling. The Top-K subset $\mathcal{K}_R \subset W$ is selected from π_R, and the rescaled cue weights $\tilde{w}_R^\ell$ preserve the total integration strength predicted for the full set W:

$$\tilde{w}_R^\ell \;=\; w_R^\ell\, \mathbf{1}[\ell \in \mathcal{K}_R] \cdot \frac{\sum_{\ell'\in W} w_R^{\ell'}}{\sum_{\ell'\in W} w_R^{\ell'}\, \mathbf{1}[\ell' \in \mathcal{K}_R]}. \tag{5}$$

The unnormalized cue strength w_R^ℓ and the preference distribution π_R^ℓ are produced by a region-cue head S_θ^r from a pooled region feature z_R^L and layer-wise region statistics ψ_R:

$$w_R^\ell \;=\; \mathrm{softplus}\big(S_\theta^r([z_R^L, \psi_R])\big)^\ell, \qquad \pi_R^\ell \;=\; w_R^\ell \,/\, \sum\nolimits_{\ell'\in W} w_R^{\ell'}, \qquad \ell \in W. \tag{6}$$

Here w_R^ℓ controls the integration strength of the residual r_R^ℓ, and π_R^ℓ ranks candidate layers for the Top-K selection in Eq. (4) and the utility-guided supervision.

Region Cue Features. The region statistics ψ_R summarize the state of candidate region cues without any class labels:

$$\psi_R = \left[\|r_R^\ell\|,\ \cos(r_R^\ell, a_R^L),\ \cos(r_R^\ell, r_R^{\ell-1})\right]_{\ell \in W}, \tag{7}$$

The coherence term is omitted for the first candidate layer. These statistics capture the magnitude of each region cue r_R^ℓ, its alignment with the semantic reference, and its cross-layer coherence (the three label-free signals shown in Sec. 2). The semantic head S_θ^a produces β_R^ℓ analogously: $\beta_R^\ell = \mathrm{softmax}(S_\theta^a([z_R^L, \psi_R]))^\ell$ for $\ell \in \mathcal{A}$, sharing the same MLP trunk with S_θ^r. Implementation details are deferred to Appendix B.

Utility-Guided Supervision. The region statistics ψ_R describe each region cue but cannot determine its recognition utility. We therefore supervise π_R with the margin-based oracle utility $u_R(\ell)$ in Eq. (3), evaluated under ReGDec's adaptive anchor: for each $\ell \in W$, we score the probe $\mathrm{Normalize}(a_R^{\mathcal{A}} + r_R^\ell)$ against the frozen text embeddings T, where $a_R^{\mathcal{A}} = \sum_{\ell \in \mathcal{A}} \beta_R^\ell \, a_R^\ell$ is the semantic readout. The resulting utility vector is converted into an online teacher $q_R = \mathrm{softmax}(u_R/\tau_u)$ with utility temperature τ_u. This teacher supervises only π_R; the semantic weights β_R receive direct supervision from the classification loss $\mathcal{L}_{\mathrm{cls}}$.

3.2 Training Objective

The fused representation $\tilde{z}_R$ is classified against the frozen text embeddings. The selector is trained with a classification loss and a utility-guided selector loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{sel}} \mathcal{L}_{\mathrm{sel}} = \mathrm{CE}_\tau\big(\mathrm{Normalize}(\tilde{z}_R) T^\top,\ y_R\big) + \lambda_{\mathrm{sel}}\, \mathrm{KL}(\,\pi_R \| \, q_R)\,. \tag{8}$$

Here, CE_τ denotes cross-entropy with CLIP-style logit scaling, and $\mathcal{L}_{\mathrm{sel}}$ is applied to the full region-cue preference distribution π_R, rather than only the selected layers in $\mathcal{K}_R$. The utility-guided loss encourages probability mass on high-utility layers while providing entropy regularization. During training, the VLM backbone and text embeddings are frozen, and only the selector S_θ is updated. Full architecture and training details are provided in Appendix B.

4 Experiments

In this section, we report the main experimental results, while additional ablation studies, analysis experiments are provided in Appendix A, and visualizations are provided in the Appendix C.

4.1 Dense Representation Evaluation

Recognition evaluation. We conduct dense-level zero-shot classification following the protocol of CLIPSelf (Wu et al., 2023b). Region representations are extracted from COCO Panoptic (Kirillov et al., 2019) via mask pooling (He et al., 2017) on Boxes, Mask Thing, and Mask Stuff. As shown in Tab. 2, gains are largest on ViT-L/14 backbones across the Top-1 metrics ($+17.1$ on CLIP, $+13.2$ on EVA-CLIP) and remain positive on their ViT-B/16 counterparts. On CLIPSelf ViT-L/14, ReGDec improves all three Top-1 metrics, with a mean gain of $+1.1$ points, indicating additional region-recognition benefits on this dense-adapted backbone. Additional results on OpenCLIP, MetaCLIP, and DFN2 in Appendix A.2 confirm the same pattern. All gains are obtained by training only the lightweight selector on 5K COCO images, with the backbone frozen. In the label-efficiency setting, 1K labeled images recover 94.6% of the gain obtained with 5K images (Appendix A.4).

Transfer to other datasets. Tab. 3 evaluates transfer across seven dense recognition benchmarks. With ViT-L/14, ReGDec achieves mean gains of 8.2 points on CLIP (Radford et al., 2021) and 6.0 points on EVA-CLIP (Sun et al., 2023). OpenCLIP, MetaCLIP, and DFN2 show gains of 4.8, 5.7, and 5.0 points, respectively (Appendix A.2). All selectors are trained on 5K COCO images and transferred without target-specific retraining. These results support cross-dataset transferability and suggest that the cross-layer distribution of useful semantic and region cues is a relatively stable property of each backbone across the evaluated datasets.

4.2 Open-Vocabulary Dense Perception Tasks

Experimental Setting. We evaluate ReGDec on two dense perception tasks: OVOD and OVSS. The source-trained selector is applied at inference without retraining the downstream detector or

Table 2: Dense-level zero-shot classification across VLM variants. We report Top-1 and Top-5 accuracy for Boxes, Mask Thing, and Mask Stuff.

Visual Encoder	VLM Variant	Boxes		Mask Thing		Mask Stuff	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
ViT-B/16	CLIP	47.9	70.1	40.4	54.7	22.8	46.1
	+ ReGDec	**51.6**	**73.6**	**41.8**	**56.3**	**27.3**	**52.0**
	EVA-CLIP	18.2	33.2	20.8	36.5	18.6	43.5
	+ ReGDec	**21.6**	**36.7**	**23.7**	**39.1**	**18.9**	**44.2**
	CLIPSelf	71.8	91.2	73.5	90.2	47.2	80.4
	+ ReGDec	**72.0**	90.9	**73.6**	90.1	46.2	77.9
ViT-L/14	CLIP	24.6	47.5	25.4	43.1	12.5	28.2
	+ ReGDec	**48.3**	**71.6**	**44.0**	**60.0**	**21.4**	**41.7**
	EVA-CLIP	56.7	78.0	57.9	78.4	21.0	42.1
	+ ReGDec	**71.4**	**89.6**	**72.6**	**87.9**	**31.1**	**57.0**
	CLIPSelf	77.2	93.3	77.9	92.4	44.9	78.5
	+ ReGDec	**78.2**	93.3	**79.3**	91.7	**45.7**	77.5

Table 3: Zero-shot evaluation of dense representations on extended datasets with ViT-L/14 backbone. The last column reports the mean improvement of ReGDec across all benchmarks.

Model	VOC	ADE150	ADE847	Context	Stuff	Object81	Objects365	Mean Δ
CLIP	41.0	9.2	4.5	18.3	3.1	7.9	9.5	–
+ **ReGDec**	**63.9**	**16.4**	**10.6**	**26.8**	**3.8**	**14.9**	**14.3**	+8.2
EVA-CLIP	84.2	39.9	27.0	40.1	4.5	12.4	38.5	–
+ **ReGDec**	**90.0**	**46.9**	**34.9**	**46.6**	**5.7**	**16.9**	**47.7**	+6.0

segmentation model. For OVOD, we build on F-ViT (Wu et al., 2023b) with frozen EVA-CLIP or CLIPSelf backbones, using the released checkpoints on OV-COCO (Lin et al., 2014) and OV-LVIS (Gupta et al., 2019). For OVSS, we follow ODISE (Xu et al., 2023b) with CLIP as the backbone, applying ReGDec to the dense features fed into the recognition module, and evaluate on ADE-150, ADE-847, and Pascal VOC. Sensitivity to box and mask perturbations is evaluated in Appendix A.5.

Open-vocabulary Object Detection. Tab. 4 reports $\mathrm{AP}_{50}^{\mathrm{N}}$ on OV-COCO and mAP_r on OV-LVIS. The selector is trained using only labeled regions from COCO base categories, without novel-category region supervision. ReGDec achieves its largest OV-COCO gain on EVA-CLIP ViT-L/14, improving novel AP by $+9.2$ ($24.2 \rightarrow 33.4$) without retraining the detector or the VLM backbone. Improvements are observed across the evaluated EVA-CLIP and CLIPSelf configurations on both benchmarks. Gains on the dense-adapted CLIPSelf backbones are smaller, suggesting that stronger dense representations leave less room for improvement through cross-layer readout.

Open-vocabulary Semantic Segmentation. As shown in Tab. 5, ReGDec provides additional recognition gains across multiple OVSS baselines, with the magnitude varying by baseline and dataset. With MaskCLIP, mIoU increases from 9.5 to 12.3 on ADE-150, from 2.8 to 3.9 on ADE-847, and from 49.5 to 61.2 on PASCAL VOC, alongside improvements in mACC. With the ODISE region-pool readout, ReGDec improves both mIoU and mACC on ADE-150 and ADE-847 across both backbones; for example, ADE-150 mIoU increases from 19.1 to 21.6 on ViT-L/14. Gains over the stronger NaCLIP baseline are more limited: ADE mIoU remains nearly unchanged, while PASCAL VOC mIoU increases from 83.0 to 84.7 and mACC improves on all three datasets. Together, these paired comparisons demonstrate that ReGDec can be integrated with different dense readout methods, while offering smaller incremental benefits in some stronger-baseline settings.

Performance and inference cost. In F-ViT with EVA-CLIP ViT-L/14 at 300 proposals, ReGDec increases mean inference latency from 187.8 to 228.7 ms/image ($|W| = 6$, latency $+21.8\%$, $\mathrm{AP}_{50}^{\mathrm{novel}}$ $+38.0\%$ relative), with 166.8 MiB additional peak allocated GPU memory. Appendix A.4 reports more details.

Table 4: **Results on OVOD.** We report $\mathrm{AP}_{50}^{\mathrm{N}}$ of the novel categories for OV-COCO and mAP_r of the rare categories for OV-LVIS. We compare our method with prior OVOD approaches and further evaluate its effect on different CLIP-based backbones.

(a) OV-COCO benchmark

Method	Backbone	$\mathrm{AP}_{50}^{\mathrm{N}}$
LP-OVOD (Pham et al., 2024)	RN50	40.5
RegionCLIP (Zhong et al., 2022)	RN50×4	39.3
CORA (Wu et al., 2023c)	RN50×4	41.7
PromptOVD (Song & Bang, 2023)	ViT-B/16	30.6
RO-ViT (Kim et al., 2023c)	ViT-L/14	33.0
CFM-ViT (Kim et al., 2023a)	ViT-L/14	34.1
EVA-CLIP (Sun et al., 2023)	ViT-B/16	18.1
+ ReGDec	ViT-B/16	**19.0**
CLIPSelf (Wu et al., 2023b)	ViT-B/16	33.6
+ ReGDec	ViT-B/16	**33.9**
EVA-CLIP	ViT-L/14	24.2
+ ReGDec	ViT-L/14	**33.4**
CLIPSelf (w/o proposals)	ViT-L/14	38.3
+ ReGDec	ViT-L/14	**38.8**
CLIPSelf (w/ proposals)	ViT-L/14	44.3
+ ReGDec	ViT-L/14	**44.8**

(b) OV-LVIS benchmark

Method	Backbone	$\mathrm{mAP}_{\mathrm{r}}$
BARON-KD (Wu et al., 2023a)	RN50	22.6
Detic (Zhou et al., 2022b)	RN50	24.9
F-VLM (Kuo et al., 2022)	RN50×4	32.8
PromptOVD (Song & Bang, 2023)	ViT-B/16	23.1
RO-ViT (Kim et al., 2023c)	ViT-B/16	28.4
CFM-ViT (Kim et al., 2023a)	ViT-L/14	33.9
DITO (Kim et al., 2023b)	ViT-L/14	**38.4**
CoDet (Ma et al., 2023)	ViT-L/14	37.0
EVA-CLIP (Sun et al., 2023)	ViT-B/16	3.3
+ ReGDec	ViT-B/16	**4.9**
CLIPSelf (Wu et al., 2023b)	ViT-B/16	25.3
+ ReGDec	ViT-B/16	**25.9**
EVA-CLIP	ViT-L/14	23.5
+ ReGDec	ViT-L/14	**23.9**
CLIPSelf	ViT-L/14	34.9
+ ReGDec	ViT-L/14	**35.4**

Table 5: **Results on OVSS.** We report mIoU and mACC on ADE-150, ADE-847, and PASCAL VOC. ReGDec uses labeled COCO source regions for selector training.

Method	VLM	ADE-150		ADE-847		PASCAL VOC	
		mIoU	mACC	mIoU	mACC	mIoU	mACC
MaskCLIP	CLIP ViT-B/16	9.5	25.0	2.8	11.0	49.5	67.5
+ ReGDec	CLIP ViT-B/16	**12.3**	**31.5**	**3.9**	**15.3**	**61.2**	**78.1**
NaCLIP	CLIP ViT-B/16	**19.0**	39.1	7.0	18.4	83.0	90.8
+ ReGDec	CLIP ViT-B/16	**19.0**	**40.4**	**7.1**	**19.6**	**84.7**	**92.0**
ODISE	CLIP ViT-B/16	20.7	31.1	4.2	6.9	83.7	95.1
+ ReGDec	CLIP ViT-B/16	**22.7**	**33.6**	**5.2**	**8.3**	83.7	**95.3**
ODISE	CLIP ViT-L/14	19.1	27.0	3.9	6.1	83.9	**95.6**
+ ReGDec	CLIP ViT-L/14	**21.6**	**30.6**	**4.9**	**7.2**	83.9	95.4

4.3 Ablation Studies

All ablations in this section are conducted on EVA-CLIP L/14 with $|W| = 6$ unless otherwise noted, evaluated on the COCO dense recognition protocol of Sec. 4.1. The reference configuration within each controlled ablation is denoted as *ours*. Additional ablations are provided in Appendix A.1.

4.3.1 Loss Components

Tab. 6 disentangles the contribution of the two loss terms in Eq. (8). Removing $\mathcal{L}_{\mathrm{cls}}$ (row 2) hurts upstream Top-1 on all three metric families and drops OVOD novel AP_{50}, confirming that ground-truth supervision anchors the selector in a class-discriminative subspace. Removing $\mathcal{L}_{\mathrm{sel}}$ (row 1) yields slightly higher upstream accuracy but lower downstream OVOD (-1.7 vs. ours), indicating that utility supervision shapes layer preferences that transfer to detection rather than overfitting to upstream classification. Within $\mathcal{L}_{\mathrm{sel}}$, our utility-based teacher improves Boxes and Mask Thing Top-1 and OVOD AP over the *Uniform* variant (row 3), with an OVOD gain of 6.4 points; Mask Stuff Top-1 is nearly unchanged (28.5 vs. 28.6). The two losses are complementary: $\mathcal{L}_{\mathrm{cls}}$ provides the discriminative anchor, and $\mathcal{L}_{\mathrm{sel}}$ with an informative teacher steers the selector toward layer choices that generalize.

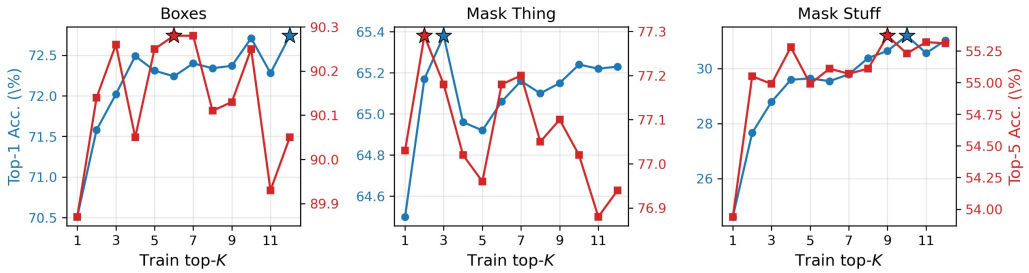


Figure 4: Train top-K sensitivity (EVA-CLIP L/14, $|W|$=12). Top-1 (blue, left) and Top-5 (red, right); stars mark per-curve best K.

Table 6: Ablation on the classification loss $\mathcal{L}_{\text{cls}}$ and the utility-guided KL loss $\mathcal{L}_{\text{sel}}$ (EVA-CLIP L/14, $|W| = 6$). The teacher distribution q_R in $\mathcal{L}_{\text{sel}}$ is either the proposed margin-based utility (*Utility*) or a non-informative *Uniform* baseline.

Row	Loss components		Top-1			OVOD
	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{sel}}$	Boxes	Mask Thing	Mask Stuff	$\text{AP}_{50}^{\text{N}}$
1	✓	–	71.3	72.8	31.0	31.7
2	–	✓	69.1	71.0	26.3	30.5
3	✓	Uniform	70.2	71.2	28.6	27.0
(ours)	✓	Utility	70.9	72.3	28.5	**33.4**

4.3.2 Selector Input Statistics ψ_R

Tab. 7 ablates the three geometric statistics fed to the selector: residual magnitude $\|r_R^\ell\|$ (*norm*), semantic redundancy with the anchor $\cos(r_R^\ell, a_R^L)$ (*dot*), and cross-layer coherence $\cos(r_R^\ell, r_R^{\ell-1})$ (*coh*). Individual statistics yield mixed results. The full set achieves the best Boxes and Mask Thing Top-1 and ties with norm-only on Mask Stuff. Relative to omitting the auxiliary statistics ψ_R while retaining the region-feature input, the full set improves the three metrics by 0.2, 0.3, and 0.2 points, respectively. Among individual cues, norm-only performs best on all three metrics. Thus, the benefit depends on the combination of statistics rather than increasing monotonically with their number.

Table 7: Ablation on selector input statistics ψ_R. *n*: $\|r_R^\ell\|$, *d*: $\cos(r_R^\ell, a_R^L)$, *c*: $\cos(r_R^\ell, r_R^{\ell-1})$.

Row	Cue Features ψ_R			Top-1		
	norm	dot	coh	Boxes	Mask	Stuff
1	–	–	–	70.7	72.0	28.3
2	✓	–	–	70.8	72.1	28.5
3	–	✓	–	70.7	71.9	27.8
4	–	–	✓	70.4	71.9	28.0
5	✓	✓	–	70.5	71.9	27.5
6	✓	–	✓	70.7	72.1	28.2
7	–	✓	✓	70.5	71.6	27.5
Ours	✓	✓	✓	**70.9**	**72.3**	**28.5**

4.3.3 Sparsity K and Anchor Configuration K_α

Train top-K sensitivity. Fig. 4 ablates the training-time sparsity K of the residual readout (Eq. (4), (5)) on Boxes, Mask Thing, and Mask Stuff. Three observations stand out. First, K=1 is consistently the worst: forcing the selector to commit to a single layer per region drops box-T@1 by 2.2, mask-T@1 by 0.9, and mask-S@1 by 6.6 relative to the saturated regime — a single residual cannot supply enough complementary information for region recognition. Second, using multiple residual layers improves over $K = 1$, although the preferred sparsity and subsequent trends vary across metrics. Third, Mask Stuff benefits from larger K, with its best Top-1 result near $K = 9$ in the plotted sweep. We use $K = 5$ as a sparse configuration for the L/14 experiments. Appendix A.3 further examines backbone-dependent effects of candidate-window size.

4.4 Selector Preference Analysis

What does the selector actually pick? We average the selector's per-layer raw weights w_R^ℓ across all regions of each type (Fig. 5). Three findings emerge. First, on *both* thing and stuff regions, layer 22 rather than the final layer 23, receives the highest mean weight, indicating that the convention of treating the final language-aligned layer as the sole readout leaves useful structure in the second-to-last layer. Second, the depth profiles differ in shape: thing regions are sharply single-peaked at layer 22, while stuff regions are broader, with elevated weights at early-mid layers ($\ell \in [13, 15]$) in addition to the peak at $\ell \in \{22, 23\}$. Third, the total predicted residual weight $\sum_\ell w_R^\ell$ is larger on stuff regions (5.3 vs. 4.1–4.3). This describes a larger weight budget for stuff regions, without directly measuring the magnitude or spatial extent of the fused residual.

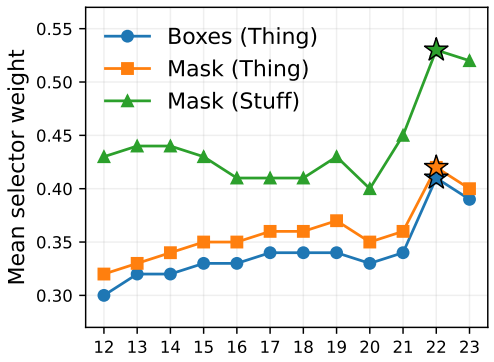


Figure 5: Selector layer-weight profile. Mean raw selector weight averaged across regions.

5 Related Work

5.1 Vision-Language Pretrained Models

CLIP (Radford et al., 2021) and subsequent VLMs (Jia et al., 2021; Ilharco et al., 2021; Sun et al., 2023; Zhai et al., 2023; Tschannen et al., 2025) enable image-level zero-shot transfer, but global alignment can limit region-level discriminability (Dong et al., 2023; Li et al., 2025). Prior work recovers local information without additional training: MaskCLIP (Zhou et al., 2022a) reuses value projections, CLIP Surgery (Li et al., 2025) modifies attention and removes redundant features, GEM (Bousselham et al., 2024) uses self-self attention, and NaCLIP (Hajimiri et al., 2025) incorporates spatial neighbourhoods. CLIPSelf (Wu et al., 2023b) instead adapts patch features through self-distillation. Building on these findings, ReGDec studies how existing region cues should be combined across depths. Beyond feature extraction and attention modification, its distinction lies in decomposing region features into CLS-aligned components and orthogonal residuals, then learning separate region-conditioned cross-layer weights for the two components. Only the selector is trained on labeled source regions, with the VLM encoders kept frozen.

5.2 Open-Vocabulary Dense Perception

VLM-based frameworks support open-vocabulary detection (Gu et al., 2021; Zang et al., 2022; Du et al., 2022; Kuo et al., 2022) and segmentation (Li et al., 2022a; Ghiasi et al., 2022; Liang et al., 2023; Xu et al., 2023c; Cho et al., 2024). Other approaches improve local representations through external spatial guidance (Zhang et al., 2025; Wang et al., 2025) or dense adaptation (Zhong et al., 2022; Rao et al., 2022). ReGDec complements downstream pipelines by improving semantic recognition of their supplied regions through learned cross-layer composition. The source-trained selector transfers across datasets without target-specific retraining, with gains depending on the backbone and existing dense representation.

6 Conclusion

We analyzed the depth-dependent behavior of region cues in frozen VLMs and observed that preferred feature depths vary across tasks and regions. These findings motivate region-adaptive cross-layer fusion of semantic and region cues. We propose **ReGDec**, a region-adaptive framework that decouples region cues from global semantics and recombines them with a lightweight utility-guided selector. With the backbone kept frozen, ReGDec consistently improves dense-level tasks.

Limitations: ReGDec relies on region cues encoded in the frozen backbone, and its gains can be limited on strong dense representations. Its benefits depend on the backbone and fusion configuration, while requiring labeled source data for selector training and additional inference time and memory.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for [Assist with translation]. We have not used generative AI tools for [Interpret results, Support qualitative and thematic data analysis, Implement methods, Design or provide feedback on research methodology or experiments, Propose or refine hypotheses], and [Clean and reformat dataset, Assist in the writing of proofs, Provide critical ingredients for proving mathematical claims, Formulate mathematical claims, Generate synthetic data sets, Help develop theoretical models or conceptual frameworks] are not applicable to this work. Additionally, we used generative AI tools for [Propose a title or keywords for a research paper, Format references, Create or modify scientific figures or images]. We have reviewed all AI-assisted work. Specifically, we checked translations of literature and text for accuracy, reviewed suggested keywords for consistency with the paper's scope, verified reference details against the cited sources, and checked the accuracy of the data presented in figures and tables. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available datasets and pretrained models.

REPRODUCIBILITY STATEMENT

The main paper and appendix provide the method formulation, training settings, evaluation protocols, and additional implementation details for reproducing our experiments. Experiments were conducted on a server equipped with eight NVIDIA Tesla V100 GPUs (32 GB each), using Python 3.8.20 and PyTorch 1.13.1 with CUDA 11.7 and cuDNN 8.5.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pp. 1931–1942. PMLR, 2021.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4123, 2024.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10995–11005, 2023.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14084–14093, 2022.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

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. Springer, 2022.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

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.

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

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Aisha Urooj Khan, Hilde Kuehne, Chuang Gan, Niels Da Vitoria Lobo, and Mubarak Shah. Weakly supervised grounding for vqa in vision-language transformers. In *European Conference on Computer Vision*, pp. 652–670. Springer, 2022.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15602–15612, 2023a.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. Detection-oriented image-text pretraining for open-vocabulary detection. 2023b.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11144–11154, 2023c.

Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6399–6408, 2019.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022b.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7061–7070, 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.

Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proceedings of the IEEE international conference on computer vision*, pp. 3040–3047, 2013.

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. *arXiv preprint arXiv:2304.07193*, 2023.

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

Congpei Qiu, Yanhao Wu, Wei Ke, Xiuxiu Bai, and Tong Zhang. Refining clip's spatial awareness: A visual-centric perspective. *arXiv preprint arXiv:2504.02328*, 2025.

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. PmLR, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18082–18091, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Hwanjun Song and Jihwan Bang. Prompt-guided transformers for end-to-end open-vocabulary object detection. *arXiv preprint arXiv:2303.14386*, 2023.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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.

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, 2023a.

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. *arXiv preprint arXiv:2310.01403*, 2023b.

Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7031–7040, 2023c.

Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 320–337. Springer, 2024.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023a.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2955–2966, June 2023b.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023c.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European conference on computer vision*, pp. 106–122. Springer, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24677–24687, 2025.

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 pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pp. 696–712. Springer, 2022a.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pp. 350–368. Springer, 2022b.

APPENDIX

A ADDITIONAL EXPERIMENTS AND ANALYSES

This appendix groups additional experiments into method ablations and readout comparisons (Sec. A.1), generalization (Sec. A.2), candidate windows and selector preferences (Sec. A.3), inference cost and label efficiency (Sec. A.4), and region-perturbation robustness (Sec. A.5). Each subsection concludes with a brief summary of its evidence.

A.1 METHOD ABLATIONS AND READOUT COMPARISONS

We examine activation, anchor, loss-weight, and candidate-window choices, then compare adaptive composition with fixed single-layer fusion and direct crop encoding.

Activation Choice for the Residual Head. A natural alternative to softplus is to apply softmax over the candidate layers, which would turn the residual weights w_R^ℓ into a probability distribution analogous to the semantic weights β_R^ℓ. The two activations differ subtly but consequentially: softplus produces *independent*, unbounded scores so each layer's injection magnitude reflects standalone need; softmax produces *competitive* scores summing to one, coupling each layer's magnitude with its relative ranking against the others. The empirical consequence is severe (Tab. 8). On CLIP ViT-B/16, softmax collapses far below the native CLIP baseline ($\overline{\Delta}_{\text{Top-1}} = -20.9$; Boxes Top-1 drops from 47.9 to 16.4), turning the residual injection into a destructive perturbation. Softplus, in contrast, improves all six accuracy metrics on CLIP ViT-B/16, with a mean Top-1 gain of $+1.3$ points over the baseline. The cause is twofold: the sum-to-one constraint of softmax forces the total residual weight to stay roughly constant regardless of how much correction the region actually requires, and it interacts poorly with our top-K truncation, which expects scores reflecting each layer's standalone usefulness rather than its share of a fixed budget. Softplus preserves this independence and lets the eventual injection magnitude be set by region-specific need rather than by competition with the discarded layers. We therefore adopt softplus as the default activation for S_θ^r.

Table 8: Ablation on the activation function of the residual head S_θ^r on CLIP ViT-B/16. Average accuracy (%); $\overline{\Delta}_{\text{Top-1}}$ averages the gain over baseline across box-/mask-Thing/Stuff Top-1. Best per column in **bold**. Results are obtained under the controlled ablation configuration used in this experiment.

Backbone	Method	Activation	Boxes		Mask Thing		Mask Stuff		$\overline{\Delta}_{\text{Top-1}}$
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	
ViT-B/16	CLIP	–	47.9	70.1	40.4	54.7	22.8	46.1	–
	w/ ReGDec	softmax	16.4	33.4	18.8	34.4	13.2	31.9	-20.9
	w/ ReGDec	softplus	**49.2**	**72.3**	**41.3**	**57.0**	**24.4**	**50.3**	$\mathbf{+1.3}$

Table 9: Ablation of K_α: *final* uses only the last layer; *average* uses equal weights over the deepest K_α layers ($K_\alpha = 2, 3, 5$); *fixed-decr.* applies fixed monotone-decreasing weights; *learned* trains the weights jointly with the selector. Top-1 accuracy (%); best per column in **bold**.

K_α	α-mode	EVA-CLIP L/14			CLIP L/14		
		Boxes	Mask Thing	Mask Stuff	Boxes	Mask Thing	Mask Stuff
baseline		56.7	57.9	21.0	24.6	25.4	12.5
1	final	70.2	71.5	28.2	43.0	40.7	**19.4**
2	average	70.2	71.8	28.3	**43.2**	40.8	19.0
3	average	70.3	71.7	28.4	43.2	**40.8**	19.0
3	fixed-decr.	**70.5**	71.6	28.8	42.7	40.5	18.9
3	learned	70.2	71.6	**28.9**	43.2	40.6	19.0
5	average	70.3	**71.9**	28.6	43.0	40.7	18.9

Anchor Configuration K_α. Tab. 9 compares semantic-anchor configurations. Across the six settings, the Top-1 ranges for Boxes, Mask Thing, and Mask Stuff are 0.3, 0.4, and 0.7 points on EVA-CLIP L/14, and 0.5, 0.3, and 0.5 points on CLIP L/14. No configuration is consistently best across backbones and metrics. Learned aggregation with K_α=3 achieves the highest Mask Stuff Top-1 on EVA-CLIP and remains within 0.5 points of the best result in every column. We use it as the default, while noting that simpler anchor constructions perform comparably in this ablation.

Selector Loss Weight λ. Fig. 6 sweeps the weight λ of $\mathcal{L}_{\text{sel}}$ across three representative teacher temperatures $\tau_u \in \{1, 3, 5\}$. The dominant trend is a near-monotonic decline as λ grows: across all three tasks, λ=1 achieves the highest Top-1 accuracy and large λ degrades performance, indicating that the utility-guided KL term should be applied as a light auxiliary signal rather than a strong objective. The teacher temperature τ_u has a smaller effect on the trend's shape but visibly shifts the operating range: $\tau_u = 1$ consistently underperforms $\tau_u = 3$ on Mask Stuff, suggesting that a sharp teacher overcommits on stuff-class regions whose useful layers are spread more evenly. Smoother teachers (τ_u=3 and τ_u=5) are nearly indistinguishable, both following the same downward trajectory. We adopt λ=1 and τ_u=3 as the default — the configuration at the intersection of the two best curves, with the smallest λ in the sweet spot of the temperature range.

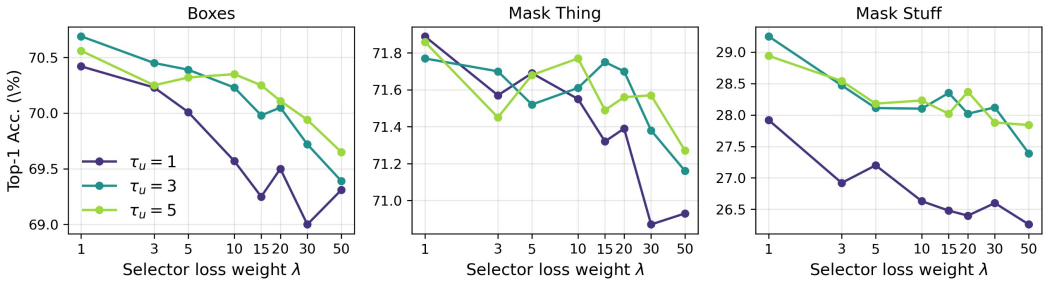


Figure 6: Sensitivity to the selector loss weight λ on EVA-CLIP L/14 ($|W| = 6$). Each panel reports Top-1 accuracy on one task as a function of λ (log scale). Three representative teacher temperatures $\tau_u \in \{1, 3, 5\}$ are shown.

Residual Candidate Set Size $|W|$. Tab. 10 ablates the size of the candidate residual layer set on EVA-CLIP L/14. The selector requires a non-trivial set: both $|W|$=3 (too restrictive) and $|W|$=24 (the entire stack) underperforms the intermediate candidate-window settings, with the latter dropping 1.7–5.7 points relative to $|W| = 12$ across the six metrics. Within the sweet spot, $|W| = 6$ and $|W| = 12$ are close on thing-class Top-1 metrics (within 0.4 points), but $|W|$=12 widens its lead substantially on stuff-class accuracy (+2.5 on mask-S@1, +1.6 on mask-S@5), suggesting that diffuse stuff regions benefit from a wider pool of candidate residuals. We adopt $|W|$=12 as the default for EVA-CLIP L/14 and use a backbone-proportional value for smaller backbones (Sec. A.3).

Table 10: Sensitivity to the residual candidate set size $|W|$ (EVA-CLIP L/14). Average accuracy (%); best per column in **bold**. Results are obtained under the controlled ablation configuration used in this experiment.

$\lvert W\rvert$	Boxes		Mask Thing		Mask Stuff	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
baseline	56.7	78.0	57.9	78.4	21.0	42.1
3	66.9	85.9	68.1	85.1	22.7	48.7
6	**70.1**	89.1	71.7	88.0	27.5	54.1
12	69.7	**89.7**	**72.1**	**88.3**	**30.0**	**55.7**
24	66.3	86.9	69.7	86.6	24.3	50.7

Single-Layer Residual Fusion vs. Adaptive Multi-Layer ReGDec. We compare fixed single-layer residual fusion with adaptive multi-layer composition. Under a fixed final-layer anchor, the readout in Eq. (4) takes the form

$$\tilde{z}_R = a_R^L + \sum_{\ell \in W} \tilde{w}_R^\ell \, r_R^\ell,$$

where $\tilde{w}_R^\ell$ are the selector's adaptive top-K weights. We collapse this to a one-hot, non-adaptive form $\tilde{z}_R^{(\ell)} = a_R^L + r_R^\ell$ for each $\ell \in \{12, \ldots, 23\}$, classifying $\mathrm{normalize}(\tilde{z}_R^{(\ell)})$ against the frozen text embeddings. The anchor is fixed at layer 23.

The fixed-anchor scan evaluates the choice of residual layer; comparison with adaptive ReGDec additionally changes the fusion configuration.

The optimal residual layer is far before the final layer. Tab. 11 shows that the residual contribution peaks at layer 19 (mask-T@1 $= 68.7$, $+8.5$ over the layer-23 residual) and layer 20 (box-T@1 $= 65.7$, $+6.9$ over layer-23). The peak is substantially before the final language-aligned layer, providing a direct, training-free experimental signal that conventional final-layer dense readouts leave usable region structure unexploited. This single-layer diagnostic mirrors the per-region oracle utility analysis of Sec. 2 and the trained-selector preference profile of Sec. 4.4: all three independent measurements identify a consistent region of the network ($\ell \in [19, 22]$) as the locus of region cues, with slight peak variations attributable to the different evaluation protocols.

Single-layer fusion recovers most of the thing-class gain but fails on stuff classes. Compared to the z_{base} baseline, the best single-layer fusion recovers approximately 69–76% of ReGDec's improvement on thing-class metrics ($+10.8$ on mask-T@1 from 57.9 to layer 19's 68.7, vs. ReGDec's $+14.2$ to 72.1). On stuff classes, however, the picture is qualitatively different: *every* single-layer fusion (including the best, layer 15 at 20.0) underperforms the baseline (21.0) on mask-S@1, while ReGDec reaches 30.0, an absolute $+9.0$ improvement. Within the tested fixed-anchor readout family, no single residual layer matches the Stuff Top-1 result of adaptive multi-layer fusion. This supports cross-layer composition over selecting a single residual layer in this setting.

Successive performance differences across readout configurations. The within-table comparison in Tab. 11 tracks successive performance differences on mask-T@1, from the naive mask-pool baseline (57.9) to ReGDec (72.1), for a total gain of 14.2 points:

- **Change of feature-extraction pipeline** ($z_{\text{base}} \to a + r$ at L=23): $+2.3$
- **Choice of residual layer** ($r_{L_{23}} \to r_{L_{19}}$): $+8.5$
- **Adaptive multi-layer fusion** (best single-layer $\to$ selector $|W|$=12): $+3.4$

These sequential comparisons describe the evaluated configurations and do not isolate independent causal contributions.

Direct Region-Crop Baseline. Using frozen OpenAI CLIP ViT-B/16 and the same 133-class text embeddings, we evaluate 55,139 COCO validation regions. Each box or masked crop is re-encoded after mean-color square padding and bicubic resizing to 224×224; masked crops replace non-region pixels with the CLIP mean.

Box cropping yields higher thing Top-1, while ReGDec yields higher Stuff Top-1 and reuses full-image features without per-crop image encoding. Results are reported in Table 12.

Summary. Within the evaluated configurations, adaptive composition improves over fixed single-layer fusion, especially on Stuff. Direct crop encoding remains stronger on thing Top-1, while ReGDec achieves higher Stuff Top-1 using shared full-image features; these comparisons do not isolate independent causal contributions.

A.2 Generalization

Across Pre-Training Recipes. To complement the three widely-used pre-training recipes reported in the main paper, we evaluate ReGDec on three additional CLIP-family variants that use different pre-training data or curation strategies: OpenCLIP (LAION-2B), MetaCLIP and DFN2. Tab. 13 reports the same dense-level zero-shot classification protocol used in Tab. 2 of the main paper. The

Table 11: Single-layer residual fusion vs. adaptive ReGDec on EVA-CLIP L/14. The top *baseline* row is the naive z_{base} pool readout (box-pool for boxes, mask-pool for masks). Each $a[L_{23}] + r[L_\ell]$ row replaces the selector's adaptive weights with a one-hot at layer ℓ (anchor fixed at layer 23). The baseline pools native dense features, whereas the $a + r$ rows use the transformed features defined in Appendix B. Same-layer reconstruction recovers these transformed features; its difference from the native baseline is not a gain from decomposition alone. The bottom row reproduces ReGDec ($|W|$=12) from Tab. 10. average accuracy (%); best single-layer per column is underlined (mask-S@1 has no meaningful single-layer best, see text); ReGDec values are in **bold**.

Setup	Boxes		Mask Thing		Mask Stuff	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
baseline (z_{base} box-/mask-pool, L= 23)	56.7	78.0	57.9	78.4	21.0	42.1
$a[L_{23}] + r[L_{12}]$	46.6	72.4	48.3	72.3	13.7	40.8
$a[L_{23}] + r[L_{13}]$	51.8	76.3	53.4	77.0	17.4	47.9
$a[L_{23}] + r[L_{14}]$	54.9	80.0	57.8	80.7	19.6	50.9
$a[L_{23}] + r[L_{15}]$	58.5	83.4	61.8	83.8	20.0	51.0
$a[L_{23}] + r[L_{16}]$	61.9	85.5	65.0	85.4	19.4	49.8
$a[L_{23}] + r[L_{17}]$	64.3	86.9	67.4	86.3	19.0	49.9
$a[L_{23}] + r[L_{18}]$	64.9	87.4	68.1	86.9	19.6	52.7
$a[L_{23}] + r[L_{19}]$	65.7	<u>87.6</u>	<u>68.7</u>	<u>87.2</u>	18.9	<u>53.4</u>
$a[L_{23}] + r[L_{20}]$	<u>65.7</u>	86.9	68.5	86.4	17.1	49.7
$a[L_{23}] + r[L_{21}]$	64.8	85.3	66.9	84.8	18.9	50.4
$a[L_{23}] + r[L_{22}]$	61.9	83.3	63.1	82.4	19.8	43.6
$a[L_{23}] + r[L_{23}]$ (final-only)	58.8	80.8	60.2	80.3	18.9	42.8
ReGDec ($\|W\|$=12)	**69.7**	**89.7**	**72.1**	**88.3**	**30.0**	**55.7**

Table 12: Direct cropping and shared-feature readout on CLIP ViT-B/16. B/T/S denote Boxes, Mask Thing, and Mask Stuff; 1/5 denote Top-1/Top-5. Crop-based B and T metrics coincide because both classify the same thing-region crop.

Readout	B@1	B@5	T@1	T@5	S@1	S@5
Box crop	54.4	72.3	54.4	72.3	20.7	53.4
Masked crop	47.0	64.3	47.0	64.3	22.8	47.8
ReGDec (Table 2)	51.6	73.6	41.8	56.3	27.3	52.0

pattern is consistent: ReGDec yields large positive gains on the L/14 backbones ($\overline{\Delta}_{\text{Top-1}}$ of $+17.6$ on OpenCLIP, $+11.3$ on MetaCLIP, and $+3.5$ on DFN2), positive but smaller gains on B/16 OpenCLIP ($+2.8$), and improves mean Top-1 across all evaluated recipes. The single sub-zero entry on DFN2 mask-Stuff@1 (-1.4) is consistent with DFN2's data-filtering training distribution, which is biased toward object-centric images and may underweight the scene-context cues required for stuff classes; even on this recipe ReGDec's mean gain across the three Top-1 metrics remains positive ($+3.5$). These results indicate that the layer-wise distribution of region cues is a property shared across the CLIP family rather than an artifact of any single pre-training recipe.

Generalization to Larger Backbones. We further validate ReGDec on substantially larger VLM backbones, specifically EVA01-CLIP ViT-G/14 and EVA02-CLIP ViT-E/14, both at 448×448 input resolution (Tab. 14). On ViT-G/14, ReGDec improves box-T@1 by $+5.8$ ($55.0 \rightarrow 60.8$), mask-T@1 by $+3.4$, and mask-S@1 by $+2.6$. The gains on ViT-E/14 are dramatic: from 18.5 to 57.3 on box-T@1 ($+38.8$), 22.7 to 59.5 on mask-T@1 ($+36.8$), and 8.0 to 27.8 on mask-S@1 ($+19.8$). The disparity between the two backbones tracks the strength of their default final-layer features for dense tasks: ViT-G/14's stronger final-layer baseline leaves less headroom for recovery, while ViT-E/14's weaker final-layer baseline benefits more from the multi-layer recombination performed by ReGDec. These results indicate that ReGDec also improves the evaluated larger backbones beyond the standard ViT-B/16 and ViT-L/14 models.

Table 13: Dense-level zero-shot classification on additional CLIP-family variants. Format follows Tab. 2 of the main paper. Top-1 and Top-5 average accuracy (%); $\overline{\Delta}_{\text{Top-1}}$ averages the gain over baseline across box-/mask-Thing/Stuff Top-1.

Visual Encoder	VLM Variant	Boxes		Mask Thing		Mask Stuff		$\overline{\Delta}_{\text{Top-1}}$
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	
ViT-B/16	OpenCLIP	49.0	71.1	41.0	54.8	30.8	56.8	–
	w/ ReGDec	**53.2**	**71.9**	**42.2**	54.7	**33.9**	56.7	+2.8
ViT-L/14	OpenCLIP	20.4	33.8	19.0	30.3	12.1	25.9	–
	w/ ReGDec	**40.0**	**62.9**	**34.7**	**52.3**	**29.6**	**53.2**	**+17.6**
	MetaCLIP	30.8	52.0	31.4	47.9	25.7	47.6	–
	w/ ReGDec	**46.0**	**68.6**	**41.7**	**58.4**	**34.2**	**59.2**	**+11.3**
	DFN2	7.3	23.4	8.3	22.6	17.1	40.1	–
	w/ ReGDec	**13.5**	**37.6**	**14.1**	**35.9**	15.7	38.1	+3.5

Table 14: Generalization of ReGDec to larger VLM backbones at 448×448 input resolution. The $|W|$ values listed are tuned per backbone rather than swept exhaustively; identifying $|W|$ from backbone properties without per-backbone tuning is left to future work. Top-1 and Top-5 accuracy (%) on classifying COCO boxes and panoptic masks (thing and stuff). $\overline{\Delta}_{\text{Top-1}}$ is the mean gain over baseline across the three Top-1 metrics.

Backbone	Method	$\lvert W\rvert$	Boxes		Mask Thing		Mask Stuff		$\overline{\Delta}_{\text{Top-1}}$
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	
ViT-G/14	EVA01-CLIP @448	–	55.0	78.4	57.1	74.7	29.1	58.1	–
	w/ ReGDec	12	**60.8**	**78.7**	**60.5**	74.7	**31.7**	**61.6**	**+3.9**
ViT-E/14	EVA02-CLIP @448	–	18.5	36.4	22.7	40.1	8.0	18.2	–
	w/ ReGDec	18	**57.3**	**79.2**	**59.5**	**75.3**	**27.8**	**59.9**	**+31.8**

Extended-Dataset Transfer on Additional CLIP Variants. To complement the main-paper transfer experiment (Tab. 3), we report the same protocol on three additional CLIP-family variants that differ in pre-training data and curation strategy: OpenCLIP (LAION-2B) (Ilharco et al., 2021), MetaCLIP (Xu et al., 2023a), and DFN2 (Fang et al., 2023). As shown in Tab. 15, ReGDec produces consistent and substantial mean gains on every variant ($+4.8$, $+5.7$, $+5.0$, respectively), and improves all seven datasets on MetaCLIP and DFN2. The single sub-baseline cell (Object81 on OpenCLIP, -0.5) is small relative to the mean improvement and does not affect the overall trend. These results reinforce the conclusion in the main paper that the layer-wise distribution of semantic and region cues exploited by ReGDec is a stable property shared across the CLIP family rather than an artifact of any single pre-training recipe.

Table 15: Zero-shot evaluation of dense representations on extended datasets with ViT-L/14 backbone, on additional CLIP-family variants. The last column reports the mean improvement of ReGDec over its corresponding baseline across all 7 benchmarks.

Model	VOC	ADE150	ADE847	Context	Stuff	Object81	Objects365	Mean Δ
OpenCLIP	52.4	11.4	4.4	20.8	2.1	12.1	8.6	–
+ **ReGDec**	**62.8**	**19.5**	**8.4**	**29.0**	**3.5**	11.6	**10.4**	+4.8
MetaCLIP	61.4	19.7	8.9	32.6	3.2	10.1	15.3	–
+ **ReGDec**	**75.2**	**27.0**	**11.3**	**41.6**	**4.1**	**13.7**	**17.9**	+5.7
DFN2	36.0	13.2	3.0	15.5	1.4	4.2	4.5	–
+ **ReGDec**	**54.8**	**14.8**	**5.2**	**23.7**	**2.9**	**5.2**	**6.1**	+5.0

Summary. Mean Top-1 gains remain positive across the evaluated additional backbones and pre-training recipes, and the source-trained selectors transfer across the tested datasets. Individual

metrics vary, so these results support transferability within the evaluated settings rather than uniform improvement on every metric.

A.3 Candidate Windows and Selector Preferences

Tab. 10 reports the $|W|$ sensitivity on EVA-CLIP L/14 with $|W| \in \{3, 6, 12\}$. Two further analyses complement those results: (i) the corresponding $|W|$ ablation on the smaller EVA-CLIP B/16 backbone, where the trend is reversed (Tab. 16); (ii) the failure mode of the largest setting $|W|$=24 on L/14 (Fig. 9).

$|W|$ on EVA-CLIP B/16. Tab. 16 reports the analogous $|W|$ sensitivity on EVA-CLIP B/16. The trend is opposite to that of L/14: the smallest setting $|W|$=3 is strongest on thing-class metrics, and thing-class Top-1 accuracy decreases as $|W|$ grows. At $|W| = 12$, the mean raw selector weights are relatively flat (Fig. 7); this observation alone does not establish redundancy between layer representations. The exception is mask-Stuff, on which moderate $|W|$ ($|W|$=6) attains the best Top-1 and Top-5 accuracy, mirroring the L/14 observation that diffuse stuff regions benefit from a wider candidate pool. Practically, we use a smaller $|W|$ for shallower backbones: $|W|$=12 for L/14 (24 transformer layers) and $|W|$=3 for B/16 (12 layers).

Table 16: Sensitivity to the residual candidate set size $|W|$ on EVA-CLIP B/16 (training epoch 15). Average accuracy (%).

$\|W\|$	Boxes		Mask Thing		Mask Stuff	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
baseline	18.2	33.2	20.8	36.5	18.6	43.5
3	**20.9**	**37.0**	**23.3**	**39.2**	18.1	43.7
6	19.5	33.8	21.8	36.9	**18.9**	**44.7**
9	19.3	33.9	21.7	36.6	18.5	44.4
12	18.6	32.9	20.7	35.5	18.5	43.5

Fig. 7 shows that on B/16, the peak layer varies with the candidate window and region type. With $|W| = 12$, the mean raw weights span approximately 0.13–0.27 and show less pronounced peaks than the L/14 profiles in Fig. 8. These profiles describe selector preferences before top-K selection and do not by themselves explain the accuracy differences.

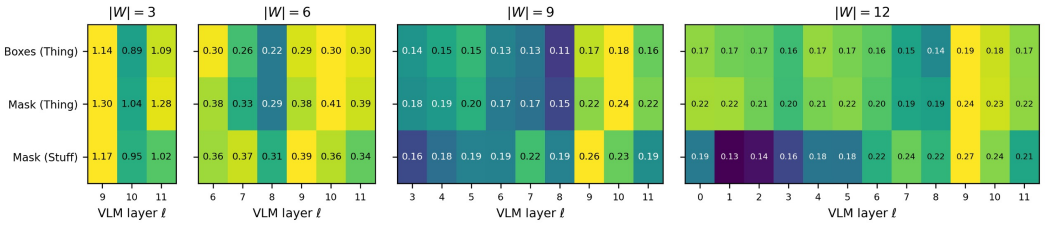


Figure 7: Selector layer-weight profile by region type on EVA-CLIP B/16. Layout matches Fig. 8. Four $|W|$ settings (3, 6, 9, 12) span a representative range. Note the flat profile at $|W|$=12, which corresponds to the accuracy collapse reported in Tab. 16.

Full-Stack Readout on EVA-CLIP L/14 ($|W| = 24$). We also show the layer profile of the $|W|$=24 configuration on EVA-CLIP L/14 (Fig. 9). This configuration underperforms $|W|$=12 on all six metrics in Tab. 10. Weights generally increase toward later layers, with local fluctuations. The broader raw-weight profile suggests less selective layer weighting, but does not establish which early-layer residuals remain after top-K selection or explain the performance decrease on its own.

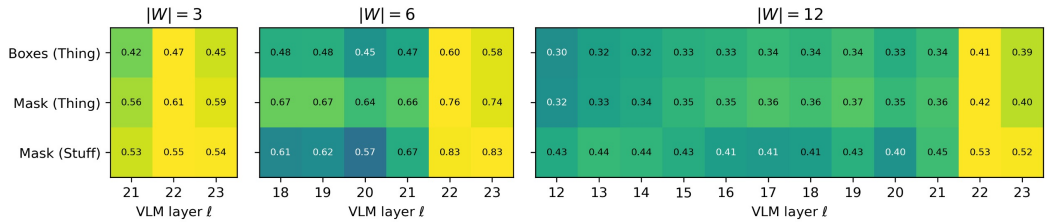


Figure 8: Selector layer-weight profile by region type, averaged across all regions of each type (EVA-CLIP L/14). Cell color is row-normalized (brightest = row maximum); text shows the absolute mean weight. Layer 22 dominates on thing regions, while stuff regions distribute mass more broadly, including elevated weights at early-mid layers when $|W|$=12.

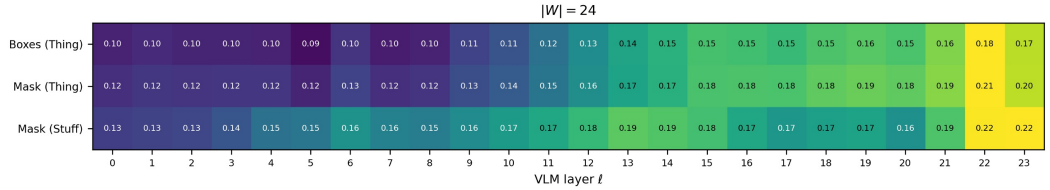


Figure 9: Selector layer-weight profile on EVA-CLIP L/14 with the full transformer stack as the candidate set ($|W|$=24). This setting underperforms $|W| = 12$ on all six metrics in Tab. 10.

Per-Layer Numerical Selector Weights. Tab. 17 reports the per-layer mean selector weights underlying Fig. 5 of the main paper. The values are raw softplus outputs of the selector's region-cue head S^r_θ averaged across all regions of each type, before the top-K sparsification in Eq. (5). The rightmost column reports the total predicted residual weight $\sum_\ell w^\ell_R$, averaged by region type. This sum is larger for stuff (5.3) than for Boxes and Mask Thing (4.1 and 4.3). It measures the weight budget, not the norm of the fused residual vector or spatial context directly.

Table 17: Selector mean per-layer weights by region type (EVA-CLIP L/14, $|W|$=12, raw softplus output averaged across all regions of each type). The largest weight per row is in **bold**, the second-largest is underlined, based on unrounded values. The rightmost column reports the total predicted residual weight $\sum_\ell w^\ell_R$, computed before rounding the per-layer weights.

Region	12	13	14	15	16	17	18	19	20	21	22	23	$\sum$
Boxes	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	**0.4**	0.4	4.1
Mask Thing	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.4	0.4	**0.4**	0.4	4.3
Mask Stuff	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.5	**0.5**	0.5	5.3

Summary. Useful candidate-window sizes and layer-weight profiles depend on the backbone and region type. Larger windows are not uniformly better, and the reported weight sums describe the selector's allocation rather than the energy or spatial extent of the fused features.

A.4 Inference Cost and Label Efficiency

Inference Latency and Memory. We benchmark F-ViT with the same EVA-CLIP ViT-L/14 detector checkpoint, COCO validation preprocessing, image order, and batch size of one on an NVIDIA V100 GPU. We compare the native single-layer model with ReGDec using six or twelve candidate layers. For each setting, we exclude 10 warmup images and measure 100 images using CUDA-synchronized wall-clock timing. Timing includes model inference but excludes data loading, evaluation, and initialization. Score-sorted RPN proposals are truncated to the specified budget before the common RoI head. Memory is peak CUDA allocated memory during inference.

Table 18 shows that latency overhead increases with both candidate-layer count and proposal budget. At 300 proposals, six and twelve layers add 40.9 and 85.9 ms/image, respectively. The corresponding

Table 18: Inference cost in F-ViT with EVA-CLIP ViT-L/14. Latency is mean ms/image; overhead is relative to the native model at the same proposal budget. Memory is peak allocated MiB.

Proposals	Readout	Latency (ms)	Overhead (%)	Memory (MiB)
50	Native	182.9	0.0	2310.9
50	ReGDec ($\|W\| = 6$)	213.6	16.8	2477.7
50	ReGDec ($\|W\| = 12$)	247.2	35.2	2646.4
100	Native	185.1	0.0	2310.9
100	ReGDec ($\|W\| = 6$)	219.6	18.6	2477.7
100	ReGDec ($\|W\| = 12$)	258.0	39.4	2646.4
300	Native	187.8	0.0	2310.9
300	ReGDec ($\|W\| = 6$)	228.7	21.8	2477.7
300	ReGDec ($\|W\| = 12$)	273.7	45.7	2646.4
1000	Native	193.9	0.0	2327.3
1000	ReGDec ($\|W\| = 6$)	249.0	28.4	2477.7
1000	ReGDec ($\|W\| = 12$)	307.4	58.5	2646.4

memory increases are 166.8 and 335.5 MiB. These measurements quantify the cost of the full inference path rather than the selector alone, and are specific to the measured detector and hardware.

Selector Supervision and Label Efficiency. We train selectors on nested subsets of 100–5,000 labeled COCO images with three seeds, using frozen EVA-CLIP ViT-L/14, $|W| = 12$, a learned three-layer anchor, and residual top-$K = 5$. Training lasts 30 epochs, with checkpoints selected by Box Top-1. We report Top-1 accuracy on Boxes, Mask Thing, and Mask Stuff; the zero-label row is the native frozen readout. Results are reported in Table 19.

Table 19: Label efficiency. Individual metrics report mean $\pm$ standard deviation across three seeds; the final column averages the three Top-1 means. Results are averaged over three seeds under the label-efficiency configuration described above.

Images	Box Top-1	Thing Top-1	Stuff Top-1	Mean
0	56.7	57.9	21.0	45.2
100	68.2±0.8	70.5±0.5	26.3±0.4	55.0
500	69.1±0.2	71.8±0.3	29.9±0.5	56.9
1000	69.8±0.2	71.9±0.1	30.4±0.6	57.4
2500	70.4±0.5	72.5±0.4	30.9±0.7	57.9
5000	70.3±0.7	72.4±0.4	31.7±0.4	58.1

With 1,000 images, the mean Top-1 gain is 12.2 points, recovering 94.6% of the 5,000-image gain (12.9 points).

Summary. In the label-efficiency setting, 1K labeled images recover most of the gain obtained with 5K images. Inference adds measurable latency and memory costs that depend on the candidate-window size and proposal budget in the measured detector setup.

A.5 Robustness to Region Perturbations

Box perturbations. For EVA-CLIP ViT-L/14 in F-ViT, baseline and ReGDec use identical perturbed COCO validation boxes. Starting from each ground-truth box, *localization jitter* translates and rescales it; *over-expansion* enlarges it to include surrounding background or neighboring content; and *under-cropping* retains a smaller sub-box inside the original box, simulating partial object coverage. Perturbation strength is chosen to place the resulting box in a target IoU bin relative to the original box. Both methods classify the source object from the same perturbed box, bypassing the RPN; this diagnostic reports region classification rather than detection AP. We report region Top-1 gains on a balanced common subset of 7,215 regions per condition, grouped by actual IoU. Results are reported in Table 20.

Table 20: Paired Box Top-1 gain (percentage points) by IoU bin.

Perturbation	$[.3, .5)$	$[.5, .7)$	$[.7, .9)$	$[.9, 1]$
Localization jitter	+0.1	+0.3	+0.8	+0.9
Over-expansion	+0.1	+0.8	+0.8	+0.8
Under-cropping	+0.4	+0.5	+0.7	+0.9

Clean-box Top-1 increases from 86.0 to 86.7. Aggregate gains are positive in all tested bins; severe perturbations can yield small negative changes for small objects.

Mask perturbations. We perturb ground-truth masks from 5,000 COCO validation images on a 14×14 grid. *Dilation* expands a mask using a square neighborhood, adding surrounding pixels; *erosion* removes boundary pixels using the same neighborhood, reducing coverage. We vary the neighborhood radius from one to six grid cells. *Neighbor merge* progressively unions the source mask with up to eight other masks from the same image, ordered by increasing centroid distance. *Split* cuts the source mask along a horizontal or vertical axis and retains one side, simulating a fragmented region. Nonempty candidates are grouped by IoU with the original mask; for each source mask, operation, and populated bin, we retain the candidate closest to the bin midpoint. The target label remains that of the source region, including for merged masks. Both methods use identical masks and frozen features; ReGDec uses residual layers $\{8, 9, 10\}$ and a learned three-layer anchor, compared with a final-layer semantic-plus-residual baseline. We report per-region Top-1 by actual mask IoU. Results are reported in Table 21.

Table 21: Mask perturbations: Top-1 (%) and paired gains (points), with region counts for each IoU bin.

Perturbation	IoU	Regions	Baseline	ReGDec	Gain
Dilation	0.3–0.5	17,813	27.7	31.9	+4.2
Dilation	0.5–0.7	7,080	27.5	33.4	+5.9
Dilation	0.7–0.9	1,367	23.4	27.9	+4.5
Dilation	0.9–1.0	15	33.3	40.0	+6.7
Erosion	0.3–0.5	2,124	42.8	50.4	+7.6
Erosion	0.5–0.7	405	40.0	44.7	+4.7
Neighbor merge	0.3–0.5	18,361	23.5	28.5	+5.0
Neighbor merge	0.5–0.7	17,353	27.1	34.1	+7.0
Neighbor merge	0.7–0.9	10,341	32.5	39.8	+7.3
Neighbor merge	0.9–1.0	4,820	33.6	41.5	+7.9
Split	0.3–0.5	24,420	34.5	40.0	+5.5
Split	0.5–0.7	30,980	35.7	42.5	+6.8
Split	0.7–0.9	23,342	38.3	45.9	+7.6
Split	0.9–1.0	30,935	37.7	45.7	+8.0

Clean-region Top-1 increases from 32.6 to 38.1 over 55,139 regions; aggregate gains remain positive in all populated perturbation bins.

Summary. Aggregate gains remain positive across the reported box and mask perturbation bins under their respective evaluation settings. This does not imply improvement for every region or object-size subgroup.

B Selector Architecture and Training Details

We provide implementation details for the selector S_θ described in Sec. 3.

Feature processing before decomposition. Let h_i^ℓ be a token at the output of transformer block ℓ. The features used in the decomposition are

$$z_i^\ell = \mathcal{N}\big(P\big(\mathrm{LN}\big(B_{\mathrm{V}}(h_i^\ell)\big)\big)\big)\,,$$

where B_{V} reuses the frozen final block with V-only attention (without QK attention mixing), LN and P are the frozen final normalization and visual projection, and $\mathcal{N}$ is token-wise L2 normalization. The

CLS token undergoes the same operations. Decomposition is performed in this shared text-embedding space, before region pooling; box features use RoIAlign and mask features use weighted averaging. Components are combined before the final region-wise L2 normalization for classification. Thus, same-layer $a_R^\ell + r_R^\ell$ reconstructs the pooled transformed features, whereas the native baseline uses the encoder's dense-feature interface. This distinction matters for interpreting Table 11.

Architecture. S_θ consists of a shared 2-layer MLP trunk followed by two parallel linear heads — a semantic head S_θ^a and a region-cue head S_θ^r. The selector input is the concatenation $[z_R^L, \psi_R] \in \mathbb{R}^{D_{\text{text}}+(3|W|-1)}$, where z_R^L is the region-pooled token at the final layer and ψ_R collects the three label-free residual statistics (residual norms, anchor alignments, and cross-layer coherences; the last has $|W|-1$ entries because layer $\ell-1$ is undefined for the smallest $\ell \in W$). The trunk produces $h_R = \mathrm{ReLU}(W_2\,\mathrm{ReLU}(W_1[z_R^L, \psi_R] + b_1) + b_2)$ with hidden dimension $H = 64$. The two heads output $\tilde{\beta}_R = W_3^a h_R \in \mathbb{R}^{|\mathcal{A}|}$ and $\tilde{w}_R = W_3^r h_R \in \mathbb{R}^{|W|}$, to which we apply softmax over $\mathcal{A}$ and element-wise softplus respectively to obtain β_R^ℓ and w_R^ℓ. Total parameters: $\sim$58K, four orders of magnitude smaller than the VLM backbone.

Region Cue Features. $z_R^L \in \mathbb{R}^{D_{\text{text}}}$ is obtained by mask-/box-pooling patch tokens within region R. The geometric statistics $\psi_R \in \mathbb{R}^{3|W|-1}$ comprise residual norms $\|r_R^\ell\|$ ($|W|$ dims), anchor alignments $\cos(r_R^\ell, a_R^L)$ ($|W|$ dims), and cross-layer coherence $\cos(r_R^\ell, r_R^{\ell-1})$ ($|W|-1$ dims). All are computed without class labels.

Training. We train with the loss in Eq. (8) using AdamW (learning rate 3×10^{-3}, batch size 64) for 30 epochs on regions from 5K COCO Panoptic training images. For OVOD, selector supervision is restricted to base-category regions. The EVA-CLIP L/14 ablation configuration uses $\lambda_{\text{sel}} = 1$, utility temperature $\tau_u = 3$, training sparsity $K = 5$, and anchor configuration $K_\alpha = 3$ with learned aggregation. This is the controlled ablation configuration, not a claim that every backbone in the broad comparison uses the same candidate-window and sparsity values. When the requested sparsity satisfies $K \geq |W|$, the implementation retains all candidate layers without top-K truncation; for positive K, the effective number of retained layers is $\min(K, |W|)$. The VLM backbone and text embeddings are kept frozen. Training takes approximately 1 hour on 8$\times$ NVIDIA V100 GPUs at 896×896 input resolution with the EVA-CLIP L/14 backbone, including data preprocessing.

C Patch Correlation Visualization



Figure 10: Patch correlation visualization on *thing* regions. **Top row**: original image with query marker. **Middle row** (EVA-CLIP): similarity map computed from EVA-CLIP L/14 final-layer patch features, after the standard projection head. The query patch is broadly correlated with most of the image, including background and unrelated objects. **Bottom row** (Ours): similarity map from the same backbone but with patch features replaced by our selector's region-adaptive multi-layer fusion. Correlation is now sharply localized to the queried object, with clear suppression of unrelated regions. No backbone parameter is modified between the two rows.

To qualitatively assess how our selector reshapes the patch representation space, we visualize the patch-to-patch cosine similarity for randomly sampled images from the COCO panoptic validation set. For each image, we select the largest thing region (according to the panoptic annotation) and place a query point at its mask centroid. We then compute, for every patch in the image, its cosine similarity to the query patch. The resulting $h \times w$ similarity map is upsampled to image resolution and overlaid on the original image with a heat colormap, as shown in Fig. 10.

(i) Diffuse correlation in baseline patches. The final-layer features of EVA-CLIP exhibit broad, image-wide correlation with the query patch. Background regions, unrelated objects, and even regions with distinctly different semantics often respond at moderate-to-high similarity. This is consistent with the global image-text alignment objective: as the network deepens, individual patch tokens are progressively shaped to match the [CLS] token's global representation, suppressing patch-specific distinctions.

(ii) Region-discriminative correlation after selector fusion. The selector-fused features concentrate similarity within the queried object region, with sharp falloff at region boundaries. Patches belonging to the same semantic region as the query light up strongly, while patches in unrelated regions are suppressed.

This visual evidence aligns with our quantitative results (Sec. 4): EVA-CLIP's final-layer patches yield diffuse correlation maps that spread well beyond the target object, whereas our selector concentrates similarity within the object boundary. The fact that the same patch-level transformation can be obtained without modifying the backbone demonstrates that the residual cues across multiple layers already contain sufficient signal that they only need to be properly selected and aggregated.