FROM FISH TO FISHING: TEACHING CLIP TO GENERALIZE THROUGH CIRCUIT-LEVEL DISTILLATION

Anonymous authors
Paper under double-blind review

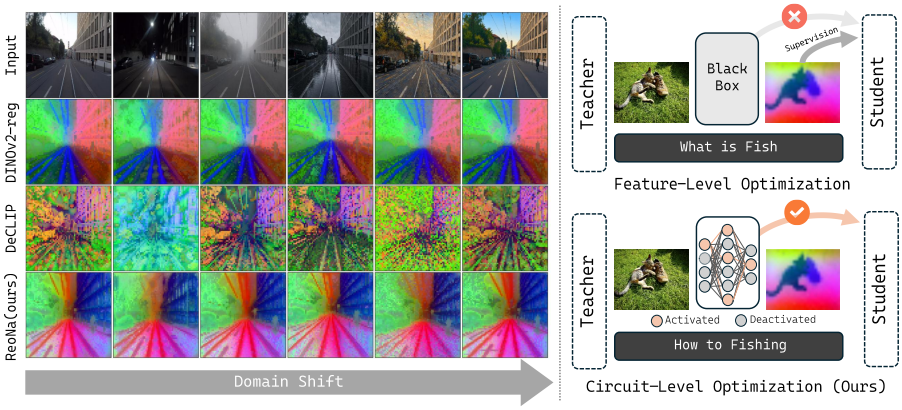

Figure 1: **Domain-robust dense features through circuit-level distillation. Left:** Under domain shifts, ReoNa produces cleaner and more consistent dense feature maps than DeCLIP. **Right:** Beyond feature-level alignment, ReoNa uses VFM cross-view consensus to guide region-conditioned channel selection within the CLIP image encoder, encouraging reuse across appearances.

ABSTRACT

CLIP aligns image content with text-defined categories, enabling dense image-text prediction. But generalizing its ability to unseen domains remains challenging. Feature-level distillation improves local semantics and spatial coherence, yet the resulting representations can still deteriorate under appearance shifts. We investigate this gap at both the representation and circuit levels, observing that the self-supervised vision foundation model (VFM) maintains more consistent dense features and greater overlap in prediction-relevant neurons than CLIP across the examined domain shifts. These observations motivate transferring not only region representations, but also the ability to reuse internal channel support across appearances. Based on that, we propose ReoNa, a novel framework that uses VFM-guided **Re**gion-t**o**-cha**N**nel distill**a**tion to teach CLIP to reuse the internal MLP channels supporting each semantic region across appearances. Extensive experiments across diverse open vocabulary domain generalized semantic segmentation (OVDG-SS) benchmarks demonstrate that ReoNa significantly improves CLIP dense representation despite domain shifts. Our code will be released.

1 INTRODUCTION

Vision–language models such as CLIP enable zero-shot image understanding through text-defined categories (Radford et al., 2021), while CLIP's image-level pretraining does not directly yield a locally discriminative dense feature map. Recent distillation-based methods strengthen CLIP's dense representations by transferring its crop-level semantics (Wu et al., 2024; Li et al., 2025b) or the spatial structure of self-supervised vision foundation models (VFMs) (Wang et al., 2025).

However, the resulting dense representations often remain sensitive to appearance changes, limiting CLIP's domain generalization capability. Fig. 1 (left) shows that prior distillation method, De-

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.

CLIP (Wang et al., 2025), produces fragmented dense features under the illustrated domain shifts, whereas its SSL-pretrained VFM teacher preserves coherent scene structure. This contrast indicates that feature-level distillation only improves local discrimination yet does not fully transfer the VFM's cross-domain robustness to CLIP. It therefore raises a central question: why does transferring locally discriminative features not necessarily transfer the ability to generalize across domains?

To understand why exploiting the VFM's dense representations does not fully transfer its robustness, we look beyond feature-level alignment and examine whether each model reuses the same prediction-relevant neurons as image appearance changes and domain shifts occur. This analysis, detailed in Sec. 2.1, reveals greater sharing of these neurons in the VFM than in CLIP across the examined shifts, highlighting reusable internal support as a key transfer target for cross-domain generalization. This insight motivates a shift *from fish to fishing*, *i.e.,* from directly aligning VFM features to teaching CLIP which internal channels are responsible for each semantic region across views. (Fig. 1, right). This results in a transfer objective that promotes reusable channel support for each semantic region across domain shifts, while maintaining consistent dense representations.

Building on this insight, we propose ReoNa, a framework that learns reusable region-to-channel support from a frozen SSL-pretrained VFM through training only on source-domain data. The VFM first aggregates region features from a source image and its spatially aligned, domain-shifted counterpart into a cross-view consensus. This consensus captures region semantics shared across domains, providing a common supervisory reference for learning domain-invariant channel support. To translate this guidance into CLIP's channel selection, we insert trainable student routers at selected encoder layers and their teacher counterparts at corresponding depths. Conditioned on the consensus, the teacher routers specify assignments over CLIP's MLP channels, while the student routers learn to predict corresponding assignments that remain consistent across views. Because the predicted assignments directly modulate CLIP's MLP activations, enforcing their consistency across views trains the image encoder to recruit the same channel support for a semantic region despite domain shifts. Moreover, inspired by masked image modeling (He et al., 2022; Zhang et al., 2025), we design a masked reconstruction objective that trains CLIP to recover the cross-view consensus from masked inputs under teacher-guided channel assignments. This reconstruction further encourages the CLIP image encoder and teacher routers to identify channel support grounded in domain-shared region semantics. Together, these objectives make reusable internal support an explicit training target. To examine whether this reusable internal support translates into stronger cross-domain generalization, we evaluate both the behavior of the learned features across domains and their performance on downstream tasks. The results consistently show that ReoNa significantly improves CLIP's cross-domain generalization capability.

Our contributions are summarized as follows:

- We identify a cross-domain sharing gap between CLIP and an SSL-pretrained vision foundation model (VFM). Across the examined domain shifts, the SSL-pretrained VFM exhibits higher dense representation similarity and greater overlap in prediction-relevant neurons, motivating distillation at both representation and circuit levels.
- We propose ReoNa, a novel framework that uses VFM cross-view consensus to guide region-to-channel distillation through learned routers. By teaching CLIP which internal MLP channels should support each semantic region, ReoNa promotes consistent region representations and reusable channel support across domains.
- Extensive experiments across diverse downstream benchmarks demonstrate that ReoNa significantly improves the generalization capability of CLIP dense feature representation under domain shifts.

2 METHODOLOGY

2.1 OBSERVATION AND MOTIVATION

As discussed in Sec. 1, feature-level distillation fails to transfer the VFM's cross-domain robustness to CLIP. To identify what CLIP should learn from the VFM to bridge this gap, we examine the cross-domain consistency of their dense representations and prediction-relevant internal support, guided by the connection between shared intermediate representations and model generaliza-

tion (Kempf et al., 2025). Specifically, we compare EVA-CLIP (Sun et al., 2023) with the self-supervised DINOv2-reg (Oquab et al., 2023). Our analysis reveals three key observations:

Observation 1: CLIP intermediate embeddings exhibit greater domain-induced separation than SSL features.

Given an image and its domain-shifted counterpart, which preserves the scene content while altering its appearance, we extract patch-level features from multiple backbone layers and jointly project the paired features into two dimensions using UMAP (McInnes et al., 2018) (Fig. 2, see Fig. 11 and Fig. 12 in appendix E for details). Across the examined layers, DINOv2-reg features from the original and domain-shifted images remain substantially intermingled, indicating that the SSL-pretrained backbone preserves a similar representation of their shared scene content despite the appearance change. In contrast, EVA-CLIP produces more clearly separated feature distributions for the two domains, which become largely disjoint at several depths. This contrast suggests that CLIP's intermediate dense representations are more sensitive to domain-specific appearance than those of the SSL-pretrained VFM.

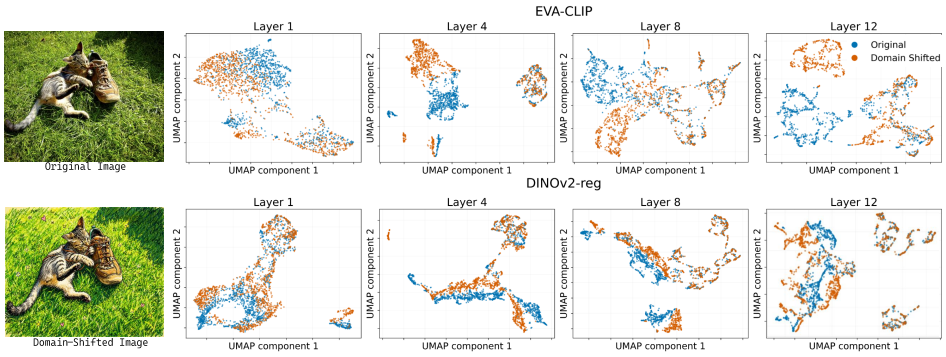

Figure 2: **Intermediate Embedding Analysis.** UMAP projections of patch-level features extracted from multiple layers of EVA-CLIP and DINOv2-reg for an image and its domain-shifted counterpart. Blue dots represent the original image, while orange dots represent the domain-shifted image.

Observation 2: SSL features maintain stronger cross-domain similarity in deeper layers.

To quantify layer-wise representation similarity beyond the UMAP visualization, we compare the ACDC (Sakaridis et al., 2021) night, fog, rain, and snow domains with their associated normal-condition reference domains. For each encoder and layer, we concatenate sampled dense patch features across images within each domain, equalize the row counts, and compute linear centered kernel alignment (CKA) (Kornblith et al., 2019). Higher CKA indicates more similar sampled-feature geometry at that layer; the computation is detailed in Appendix B.1.

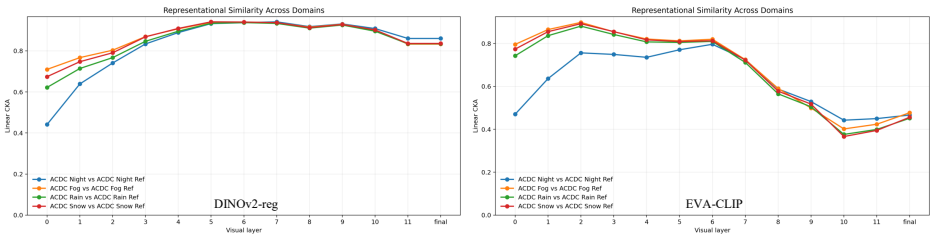

Figure 3: **Layer-wise Cross-Domain Representation Similarity.** Linear CKA between domain-level matrices of sampled dense patch features from ACDC adverse-weather and reference domains. DINOv2-reg maintains high similarity across depth, whereas EVA-CLIP similarity decreases markedly in later layers.

As shown in Fig. 3, the early EVA-CLIP layers retain relatively high similarity across weather conditions, suggesting that they still encode shared domain-invariant semantics common to each image pair. However, CKA decreases substantially toward the final visual representation, indicating that deeper CLIP features become increasingly affected by weather, illumination, and imaging style. In contrast, DINOv2-reg sustains consistently high CKA through most of the backbone. Its dense features therefore exhibit stronger and more persistent cross-domain representation sharing than those of EVA-CLIP, especially in deeper layers.

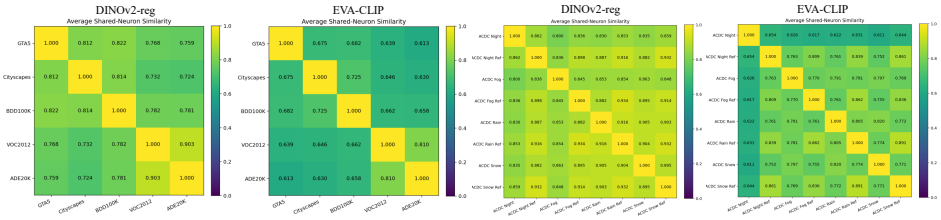

Figure 4: **Cross-Domain Circuit Similarity.** Average Jaccard overlap between sets of prediction-relevant neurons across dataset and adverse-weather domain pairs. DINOv2-reg exhibits consistently higher shared-neuron similarity than EVA-CLIP.

Observation 3: An SSL-pretrained VFM exhibits stronger cross-domain circuit similarity than CLIP.

To examine whether cross-domain representation sharing is also reflected in the model's internal computational support, we compare the prediction-relevant neurons identified for matched dense semantic predictions across domains. We measure the overlap between these neuron sets using Jaccard similarity, where a higher score indicates that a model relies on more shared internal units across a domain pair. Fig. 4 summarizes the average shared-neuron similarity under both cross-dataset shifts and ACDC adverse-weather shifts. We defer the full attribution and neuron-selection protocol to the appendix B.2.

Across representative urban dataset pairs, DINOv2-reg attains shared-neuron similarities of 0.812–0.822, compared with 0.675–0.725 for EVA-CLIP. The same pattern persists across ACDC weather/reference pairs, where DINOv2-reg reaches 0.845–0.916 while EVA-CLIP ranges from 0.654 to 0.805. Together with the CKA analysis, these results indicate that the domain-shared dense representations of the SSL backbone are accompanied by more stable prediction-relevant neuron support, whereas CLIP relies on internal support that varies more strongly with domain appearance.

Motivation. Our observations show that the SSL-pretrained VFM maintains more consistent intermediate representations across layers and greater sharing of prediction-relevant neurons than CLIP under domain shifts. Existing feature-level distillation methods, however, focus on dense-output alignment but leave these intermediate and circuit-level properties insufficiently constrained across domains, limiting the transfer of the VFM's robustness. We therefore aim to *distill multi-layer representation consistency and circuit-level sharing* from the VFM into CLIP, so that each semantic region retains stable intermediate features and reusable internal channel support across appearance changes.

2.2 Proposed Method

We propose ReoNa, a teacher–student framework that improves the domain invariance of a CLIP image encoder through training only on source-domain data. As shown in Fig. 5, we aggregate a frozen SSL-pretrained VFM's region features from an image and its appearance-augmented version into a cross-view consensus that serves as a shared semantic target. Using this consensus, the framework trains CLIP to learn domain-invariant region representations and region-to-channel assignments that encourage the reuse of prediction-relevant neurons across appearance changes.

Constructing Cross-view Consensus. We first construct a region-level teacher target that retains the semantic evidence shared by two appearances of the same scene. For each source image $x_i \in \mathbb{R}^{H\times W\times 3}$, we apply a photometric transformation $\mathcal{A}$ to obtain a shifted view $\hat{x}_i = \mathcal{A}(x_i)$. The clean and shifted images are both passed through the frozen SSL-pretrained VFM Φ_{vfm}, producing dense

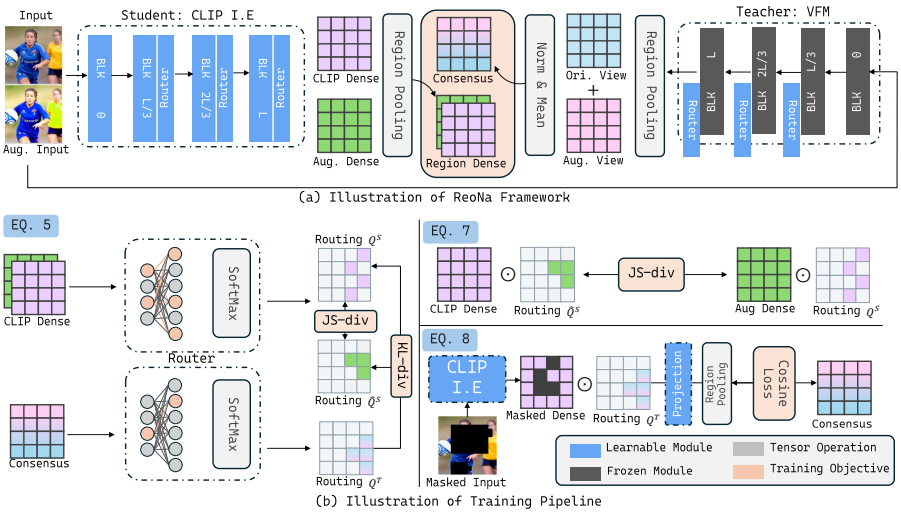

Figure 5: **Overview of ReoNa. (a)** Model architecture, with routers at selected depths and cross-view region consensus from a frozen VFM. **(b)** Training combines region-to-channel distillation (left), cross-view routing exchange for prediction consistency (upper right), and teacher-guided masked reconstruction of the VFM consensus (lower right).

feature maps f_i and $\hat{f}_i$:

$$f_i = \Phi_{\text{vfm}}(x_i), \qquad \hat{f}_i = \Phi_{\text{vfm}}(\hat{x}_i), \qquad f_i, \hat{f}_i \in \mathbb{R}^{D_f \times H_f \times W_f}, \tag{1}$$

where D_f and $H_f \times W_f$ denote the teacher feature dimension and spatial resolution, respectively.

Because $\mathcal{A}$ preserves image geometry, the same semantic regions remain spatially aligned across the two views. We derive R_i valid regions from the source annotation y_i and resize their binary masks to the teacher feature grid, yielding $M_i \in \{0,1\}^{R_i \times H_f \times W_f}$. For region r, we pool the dense features within its mask as

$$F_{i,r} = \frac{\sum_p M_{i,r,p} f_{i,p}}{\sum_p M_{i,r,p} + \epsilon}, \qquad \hat{F}_{i,r} = \frac{\sum_p M_{i,r,p} \hat{f}_{i,p}}{\sum_p M_{i,r,p} + \epsilon}, \qquad F_{i,r}, \hat{F}_{i,r} \in \mathbb{R}^{D_f}, \tag{2}$$

where p indexes the $H_f W_f$ spatial locations and ϵ is a small constant for numerical stability. Stacking all regions gives $F_i, \hat{F}_i \in \mathbb{R}^{R_i \times D_f}$.

We then define the cross-view consensus target by averaging the independently normalized region representations:

$$T_i^{\text{cons}} = \text{Norm}\left(\frac{\text{Norm}(F_i) + \text{Norm}(\hat{F}_i)}{2}\right), \qquad T_i^{\text{cons}} \in \mathbb{R}^{R_i \times D_f}, \tag{3}$$

where $\text{Norm}(\cdot)$ denotes row-wise ℓ_2 normalization. The resulting T_i^{cons} is used as the common teacher target for channel routing and representation reconstruction.

Routing. We then use the VFM consensus to supervise which CLIP channels support each semantic region. Let the CLIP image encoder Φ_{clip} contain L Transformer blocks. We introduce routing at block indices $\mathcal{S} = \{\lfloor L/3 \rfloor, \lfloor 2L/3 \rfloor, L\}$, and denote the number of MLP expansion channels at block $l \in \mathcal{S}$ by C_l. At each selected depth, we insert a student router $\mathcal{R}_S^l$ into the student network and add a VFM-guided teacher router $\mathcal{R}_T^l$ at the corresponding depth of the teacher branch.

The teacher router takes the consensus features $T_i^{\text{cons}} \in \mathbb{R}^{R_i \times D_f}$ as its region-level input. For the student router, we denote the intermediate patch-token representations output by CLIP block l as $U_i^l \in \mathbb{R}^{N_l \times D_l}$, where N_l is the number of spatial patch tokens and D_l is their feature dimension. We resize each region mask in M_i to the corresponding token grid and average-pool the covered rows of U_i^l. Stacking the resulting vectors gives the student input $A_i^l \in \mathbb{R}^{R_i \times D_l}$.

Each router maps its region features to logits over the C_l MLP channels. A row-wise softmax produces the teacher and student routing distributions

$$Q_i^{T,l} = \mathrm{softmax}_{C_l}\big(\mathcal{R}_T^l(T_i^{\mathrm{cons}})\big), \qquad Q_i^{S,l} = \mathrm{softmax}_{C_l}\big(\mathcal{R}_S^l(A_i^l)\big), \tag{4}$$

where $Q_i^{T,l}, Q_i^{S,l} \in \mathbb{R}^{R_i \times C_l}$, and $\mathrm{softmax}_{C_l}$ is applied row-wise over the MLP-channel dimension. Each MLP hidden channel corresponds to a neuron, so these routing distributions specify how strongly each neuron should support a region's prediction. Thus, $Q_i^{T,l}$ provides a VFM-guided channel-support target, whereas $Q_i^{S,l}$ estimates the corresponding assignment from CLIP features alone. Each row directly parameterizes a gate over the C_l hidden channels of its associated MLP, making prediction-relevant neuron support for each region an explicit object of supervision.

Training Procedure. An unrouted student pass over x_i and $\hat{x}_i$ provides the block features A_i^l and $\hat{A}_i^l$, from which the student routers predict $Q_i^{S,l}$ and $\hat{Q}_i^{S,l} \in \mathbb{R}^{R_i \times C_l}$. Both predictions are supervised by the same consensus-derived teacher matrix $Q_i^{T,l}$.

Route. We distill the VFM-guided region-to-channel assignment by aligning the routing matrix predicted from each student view with the stopped-gradient teacher matrix. We additionally constrain the two student routing matrices to remain consistent across the appearance change:

$$\mathcal{L}_{\mathrm{route}} = \mathbb{E}_i \frac{1}{|\mathcal{S}|} \sum_{l \in \mathcal{S}} \left[\frac{1}{2} \mathrm{KL}\Big(Q_i^{S,l} \,\|\, \mathrm{sg}(Q_i^{T,l})\Big) + \frac{1}{2} \mathrm{KL}\Big(\hat{Q}_i^{S,l} \,\|\, \mathrm{sg}(Q_i^{T,l})\Big) + \beta\, \mathrm{JS}\Big(Q_i^{S,l}, \hat{Q}_i^{S,l}\Big) \right], \tag{5}$$

where $\mathbb{E}_i$ denotes the expectation over source training samples, and KL and JS denote row-wise Kullback–Leibler and Jensen–Shannon divergences, respectively. Simultaneously, we update $\mathcal{R}_T^l$ using the JS divergence between F_i and $\hat{F}_i$. To apply a routing matrix to the student, let $\bar{M}_i^l \in [0,1]^{N_l \times R_i}$ be the region-membership matrix on the token grid of block l, normalized across regions. The corresponding mean-one channel gate is

$$\mathcal{G}_i^l(Q) = \mathbf{1}_{N_l \times C_l} + \alpha \bar{M}_i^l \left(C_l Q - \mathbf{1}_{R_i \times C_l}\right) \in \mathbb{R}^{N_l \times C_l}, \tag{6}$$

where $Q \in \mathbb{R}^{R_i \times C_l}$ is a routing matrix. This gives $G_i^l = \mathcal{G}_i^l(Q_i^{S,l})$ for x_i and $\hat{G}_i^l = \mathcal{G}_i^l(\hat{Q}_i^{S,l})$ for $\hat{x}_i$. Each gate is applied to the normalized MLP hidden activation before its output projection.

Swap. To require the learned channel support to transfer across appearances, we cross-apply the student routing matrices between the two aligned views: x_i is processed using $\hat{G}_i^l$, while $\hat{x}_i$ is processed using G_i^l. Let $P_i^{x \leftarrow \hat{x}}, P_i^{\hat{x} \leftarrow x} \in [0,1]^{R_i \times N_b}$ denote the resulting row-wise routed predictions over the N_b base classes. We enforce their agreement through

$$\mathcal{L}_{\mathrm{swap}} = \mathbb{E}_i\ \mathrm{JS}\big(P_i^{x \leftarrow \hat{x}}, P_i^{\hat{x} \leftarrow x}\big). \tag{7}$$

Agreement after this exchange encourages the learned channel support to remain functional across appearances, rather than merely matching the teacher within each view.

Masked Reconstruction. We apply a random patch mask to x_i and feed the masked image through the CLIP image encoder. During this forward pass, the teacher-guided gate $G_i^{T,l} = \mathcal{G}_i^l(Q_i^{T,l})$ modulates the MLP-channel contributions of the patch-token representations $U_{i,\mathrm{mask}}^l$ at each selected block l. We project the resulting masked dense features into the VFM feature space and pool them over all regions, yielding $S_i^{\mathrm{mask}} \in \mathbb{R}^{R_i \times D_f}$. To reconstruct the cross-view consensus, we optimize the row-wise cosine loss

$$\mathcal{L}_{\mathrm{recon}} = 1 - \mathbb{E}_i \frac{1}{R_i} \left\langle \mathrm{Norm}(S_i^{\mathrm{mask}}), \mathrm{sg}(T_i^{\mathrm{cons}}) \right\rangle_F, \tag{8}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. $\mathcal{L}_{\mathrm{recon}}$ simultaneously updates the CLIP image encoder and the teacher routers, encouraging them to learn channel support for domain-invariant region semantics.

The resulting domain-sharing objective is

$$\mathcal{L} = \lambda_{\mathrm{route}} \mathcal{L}_{\mathrm{route}} + \lambda_{\mathrm{swap}} \mathcal{L}_{\mathrm{swap}} + \lambda_{\mathrm{recon}} \mathcal{L}_{\mathrm{recon}}, \tag{9}$$

Together, these terms make region representations consistent across views while encouraging corresponding regions to reuse the same CLIP channel support.

Table 1: **Zero-shot performance.** Results are reported in mIoU.

	Dv-19					Dv-58				
Method	CS-19	ACDC-19	BDD-19	Mapi-19	Avg.	ACDC-41	BDD-41	Mapi-30	RW-10	Avg.
EVA-CLIP (Sun et al., 2023)	26.4	22.9	25.8	28.8	26.0	24.4	32.1	24.0	21.1	25.4
CLIPSelf (Wu et al., 2024)	27.0	21.8	26.0	30.6	26.4	22.2	28.1	19.2	22.2	22.9
ATAS (Yeo et al., 2025)	28.1	25.6	26.2	33.7	28.4	25.8	28.7	21.2	23.4	24.8
DenseVLM (Li et al., 2025b)	29.5	24.1	28.3	34.8	29.2	30.0	34.0	26.3	27.6	29.5
DeCLIP (Wang et al., 2025)	32.5	30.0	30.9	37.1	32.6	32.4	39.6	33.1	25.7	32.7
ReoNa (ours)	**44.2**	**40.7**	**43.8**	**47.2**	**44.0**	**42.4**	**48.8**	**37.5**	**28.8**	**39.4**

3 EXPERIMENTS

We use open-vocabulary domain-generalized semantic segmentation (OVDG-SS) benchmarks (Zhao et al., 2026) as a testbed for evaluating the cross-domain generalization of CLIP's dense representations. In this setting, a model is trained only on labeled source-domain images and evaluated on unseen domains using a text-provided vocabulary that includes both source classes and novel classes absent from the source annotations. These benchmarks therefore assess whether its dense features remain locally discriminative despite the simultaneous domain and vocabulary shifts. We evaluate this capability through both direct visual–text matching and finetuned decoders. Implementation details are provided in Appendix A.

3.1 DATASETS AND EVALUATION

Following Zhao et al. (2026), we evaluate on Dv-19 and Dv-58 under standard and expanded vocabularies, respectively. Dv-19 comprises Cityscapes-19 (CS-19) (Cordts et al., 2016), ACDC-19 (Sakaridis et al., 2021), BDD-19 (Yu et al., 2020), and Mapi-19, sharing the Cityscapes 19-class vocabulary across diverse weather and illumination conditions. Mapi-19 uses remapped Mapillary (Neuhold et al., 2017) validation annotations.

Dv-58 comprises ACDC-41 and BDD-41, extended to 41 categories through diffusion-based object inpainting and manual refinement; Mapi-30, grouping Mapillary labels into 30 coarse classes; and RW-10, grouping ROADWork (Ghosh et al., 2025) labels into 10 construction-related classes. We report mean intersection-over-union (mIoU) per test set and the unweighted average across test sets within each benchmark setting.

3.2 BENCHMARK RESULTS

Zero-shot OVDG-SS based on VLM features. We compare ReoNa with CLIP-based baselines by directly matching dense visual features to class-text embeddings, without training an additional segmentation decoder. As shown in Table 1, ReoNa achieves 44.0 mIoU on Dv-19 and 39.4 mIoU on Dv-58 on average, exceeding DeCLIP, the strongest baseline on average, by 11.4 and 6.7 percentage points, respectively. These results show that ReoNa substantially strengthens the cross-domain generalization of CLIP's dense representations under direct visual–text matching.

Table 2: **Synthetic-to-Real performance** (mIoU). ViT-B/16 baselines follow Zhao et al. (2026).

	Dv-19					Dv-58				
Method	CS-19	ACDC-19	BDD-19	Mapi-19	Avg.	ACDC-41	BDD-41	Mapi-30	RW-10	Avg.
CAT-Seg (Cho et al., 2024)	43.7	37.6	45.4	48.8	43.9	51.2	56.2	39.3	35.7	45.6
ESC-Net (Lee et al., 2025)	42.2	35.9	44.6	48.1	42.7	49.4	52.4	37.9	33.3	43.3
Mask-Adapter (Li et al., 2025a)	46.0	36.4	45.7	50.1	44.6	50.1	49.2	38.1	32.7	42.5
CAT-Seg+AdvStyle (Zhong et al., 2022)	44.9	38.2	43.8	51.1	44.5	49.2	54.2	39.1	34.3	44.2
CAT-Seg+DGInStyle (Jia et al., 2024)	44.4	39.1	48.1	51.4	45.8	50.7	55.9	39.6	34.0	45.1
S^2-Corr (Zhao et al., 2026)	49.3	40.4	48.6	54.5	48.2	51.3	56.4	42.4	36.7	46.7
CLIPSelf (Wu et al., 2024)+S^2-Corr	49.5	42.0	50.3	55.5	49.3	51.1	57.8	41.5	35.4	46.5
DenseVLM (Li et al., 2025b)+S^2-Corr	50.3	43.3	52.1	56.9	50.7	52.0	59.6	44.7	37.3	48.4
DeCLIP (Wang et al., 2025)+S^2-Corr	51.2	44.6	52.5	58.0	51.6	50.1	58.7	45.2	36.2	47.6
ReoNa+S^2-Corr (ours)	**53.0**	**46.8**	**53.8**	**59.5**	**53.3**	**55.2**	**62.5**	**46.7**	**37.8**	**50.6**

OVDG-SS via fine-tuned decoders. Following Zhao et al. (2026), we pair ReoNa with the S^2-Corr state-space-model decoder and fine-tune the decoder on GTAV-7 (Richter et al., 2016) for **Synthetic-to-Real** and Cityscapes-7 (Cordts et al., 2016) for **Real-to-Real**, comparing against existing segmentation methods and CLIP dense-feature refinement baselines. Under Synthetic-to-Real (Table 2), ReoNa+S^2-Corr achieves 53.3 mIoU on Dv-19 and 50.6 mIoU on Dv-58 on average, exceeding the strongest baseline for each benchmark by 1.7 and 2.2 percentage points, respectively. Under Real-to-Real (Table 3), it reaches 55.9 mIoU and 53.0 mIoU, respectively, attaining the highest average mIoU on both benchmarks. These results show that the cross-domain benefits of ReoNa's dense representations extend to segmentation with fine-tuned decoders. Performance of additional decoder is presented in Appendix C.4.

Table 3: **Real-to-Real performance** (mIoU). ViT-B/16 baselines follow Zhao et al. (2026).

	Dv-19				Dv-58				
Method	ACDC-19	BDD-19	Mapi-19	Avg.	ACDC-41	BDD-41	Mapi-30	RW-10	Avg.
CAT-Seg (Cho et al., 2024)	38.9	44.0	47.6	43.5	47.6	51.6	38.2	36.5	43.5
ESC-Net (Lee et al., 2025)	40.4	43.6	45.9	43.3	47.3	49.2	40.1	34.5	42.8
Mask-Adapter (Li et al., 2025a)	38.9	47.7	49.9	45.5	48.3	51.2	39.8	36.1	43.9
CAT-Seg+AdvStyle (Zhong et al., 2022)	37.2	43.2	44.6	41.7	48.0	47.9	36.0	36.5	42.1
CAT-Seg+DGInStyle (Jia et al., 2024)	42.5	49.8	49.4	47.2	46.1	53.7	36.8	36.0	43.2
S^2-Corr (Zhao et al., 2026)	44.6	50.1	56.2	50.3	52.3	58.6	42.0	38.6	47.9
CLIPSelf (Wu et al., 2024)+S^2-Corr	49.0	53.0	60.3	54.1	55.9	60.3	44.3	39.7	50.1
DenseVLM (Li et al., 2025b)+S^2-Corr	50.1	53.1	60.2	54.5	57.0	62.1	46.5	42.2	52.0
DeCLIP (Wang et al., 2025)+S^2-Corr	51.2	53.5	61.0	55.2	53.4	58.7	46.7	41.5	50.1
ReoNa+S^2-Corr (ours)	**51.9**	**54.0**	**61.8**	**55.9**	**57.4**	**63.4**	**47.6**	**43.7**	**53.0**

4 Analysis of ReoNa

Effect of Each Component. We evaluate the contribution of each component by progressively adding region-to-channel routing, cross-view routing exchange (Swap), and masked reconstruction to a feature-alignment baseline (Table 4). Region-to-channel routing provides the largest gains, improving average performance by 8.7 mIoU on Dv-19 and 5.9 mIoU on Dv-58, supporting the value of region-conditioned channel supervision beyond feature alignment. Swap adds a further 1.3 mIoU and 0.3 mIoU, respectively, while masked reconstruction contributes another 1.4 mIoU and 0.5 mIoU, bringing the full model to 44.0 mIoU and 39.4 mIoU. Each successive addition improves both benchmarks, supporting the combined use of channel supervision, cross-view exchange, and masked reconstruction for hidden semantic cues.

Table 4: Cumulative component ablation of ReoNa (average mIoU).

Component	Dv-19	Dv-58
Feature alignment (DeCLIP)	32.6	32.7
Region-to-channel routing†	41.3	38.6
+ Swap	42.6	38.9
+ Mask recon.‡ (ReoNa)	**44.0**	**39.4**

† Region-to-channel routing uses the ReoNa framework with $\mathcal{L}_{\text{route}}$ and unmasked $\mathcal{L}_{\text{recon}}$. ‡ The masked-reconstruction ablation computes the reconstruction loss without masking any regions.

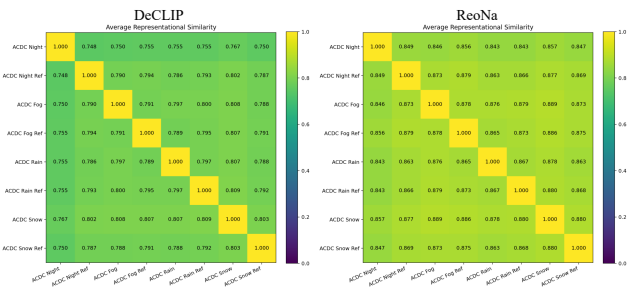

Figure 6: **Shared-neuron similarity.** Average Jaccard overlap across ACDC domain pairs for DeCLIP and ReoNa.

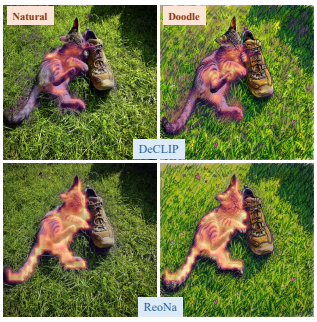

Figure 7: **Fixed-channel responses** to "cat" in natural and doodle views. Rows show DeCLIP and ReoNa.

Analysis of Shared Neurons. Fig. 6 compares DeCLIP and ReoNa using the same Jaccard-based protocol as Fig. 4. Across ACDC weather and reference domains, ReoNa achieves consistently higher shared-neuron similarity than the feature-level distillation baseline DeCLIP, indicating greater reuse of prediction-relevant neurons under appearance shifts. Fig. 7 examines this reuse through individual concept responses: independently for each model, we select the channel with the strongest response to "cat" in the natural image and visualize that same channel in its doodle counterpart. For ReoNa, the fixed channel retains a strong response over the cat region after the shift, whereas DeCLIP's response becomes weaker and less coherent. Together, the higher shared-neuron overlap and preserved concept responses support the effectiveness of region-to-channel distillation in learning reusable internal support across appearances.

5 Related Work

5.1 Dense Representations of CLIP

CLIP's image-level contrastive pretraining provides limited supervision for locally discriminative features, motivating methods that improve its dense representations (Radford et al., 2021; Jin et al., 2025). OVSeg (Liang et al., 2023) adapts CLIP to masked regions using region–text pairs and mask prompt tuning. FC-CLIP (Yu et al., 2023) shares a frozen convolutional CLIP backbone between mask prediction and open-vocabulary classification, while CAT-Seg (Cho et al., 2024) aggregates dense image–text similarity scores and fine-tunes CLIP for segmentation.

Recent approaches concentrate on fine-tuning CLIP's image encoder to enhance local region features. CLIPSelf (Wu et al., 2024) aligns pooled region features with CLIP embeddings of corresponding image crops through self-distillation. DenseVLM (Li et al., 2025b) reduces foreground bias in region–language alignment, ATAS (Yeo et al., 2025) uses self-distillation across representation levels to improve semantic coherence and fine-grained alignment. DeCLIP (Wang et al., 2025) decouples content and context learning, combining crop-level CLIP supervision with DINO-guided spatial correlations. These methods refine CLIP's local features but suffer from domain shifts.

5.2 Domain Generalization of CLIP

Various studies have examined how pretraining data shape CLIP's generalization. Fang et al. (2022) identify training-distribution diversity as the primary driver of robustness gains in their experiments. Mayilvahanan et al. (2024) shows that high train–test similarity alone cannot explain CLIP's out-of-distribution performance, while Mayilvahanan et al. (2025) finds that withholding entire visual styles substantially reduces transfer to those styles, highlighting the importance of pretraining domain coverage. Xue et al. (2024) explains how rich captions can improve robustness through intra-class contrasting and inter-class feature sharing. At the mechanistic level, Kempf et al. (2025) link domain and compositional generalization to shared intermediate representations and prediction-relevant circuits, showing improved alignment of final embeddings can coexist with poor transfer.

Complementary methods use CLIP to improve domain-generalized classification. RISE (Huang et al., 2023) regularizes a student's visual representations with CLIP text embeddings, while CLIP-CEIL (Yu et al., 2024) uses an adapter to reduce cross-domain variation in visual feature channels while preserving class discrimination and image–text alignment. For dense representations, S^2-Corr (Zhao et al., 2026) addresses domain shifts via state-space-model (Liu et al., 2024) decoder.

6 Conclusion

We presented ReoNa, a framework for improving the cross-domain generalization of CLIP's dense representations. Our analysis showed that, across the examined domain shifts, an SSL-pretrained VFM exhibits more consistent representations and greater overlap in prediction-relevant neurons than CLIP. Guided by these observations, ReoNa uses VFM cross-view consensus to supervise region-conditioned MLP-channel routing, jointly encouraging consistent representations and channel reuse across appearances. Comprehensive evaluations show that this supervision yields more robust CLIP dense representations in unseen domains, demonstrating the value of transferring reusable internal channel support.

AI USE STATEMENT

In this work, we used generative AI tools to aid writing and coding. We did not use generative AI tools to develop the research idea, conduct the literature review, collect data, or perform analysis. LLM-generated code was verified and tested for correctness by two authors, and all AI-polished writing was reviewed by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main paper specifies the model formulation, training objectives, datasets, and evaluation protocols. The appendix provides the model architecture, optimization hyperparameters, training details, technical definitions of CKA and shared-neuron similarity, ablations, per-class results, and computational cost. All datasets are publicly available. We provide the core implementation code in supplementary files, full code and configuration files will be made publicly available.

REFERENCES

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, 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 *CVPR*, pp. 3213–3223, 2016.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pp. 226–231, 1996.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (CLIP). In *ICML*, pp. 6216–6234, 2022.

Anurag Ghosh, Shen Zheng, Robert Tamburo, Khiem Vuong, Juan Alvarez-Padilla, Hailiang Zhu, Michael Cardei, Nicholas Dunn, Christoph Mertz, and Srinivasa G Narasimhan. Roadwork: A dataset and benchmark for learning to recognize, observe, analyze and drive through work zones. In *ICCV*, pp. 6132–6142, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling CLIP with language guidance. In *ICCV*, pp. 11685–11695, 2023.

Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. DGInStyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *ECCV*, 2024.

Shuo Jin, Siyue Yu, Bingfeng Zhang, Mingjie Sun, Yi Dong, and Jimin Xiao. Feature purification matters: Suppressing outlier propagation for training-free open-vocabulary semantic segmentation. In *ICCV*, pp. 20291–20300, 2025.

Elias Kempf, Simon Schrodi, Max Argus, and Thomas Brox. When and how does clip enable domain and compositional generalization? In *ICML*, pp. 29706–29734, 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 *ICCV*, pp. 4015–4026, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, pp. 3519–3529, 2019.

Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective SAM combination for open-vocabulary semantic segmentation. In *CVPR*, pp. 26081–26090, 2025.

Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In *CVPR*, pp. 14998–15008, 2025a.

Yunheng Li, Yuxuan Li, Quansheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased region-language alignment for open-vocabulary dense prediction. In *ICCV*, pp. 23795–23805, 2025b.

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 *CVPR*, pp. 7061–7070, 2023.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *NeurIPS*, pp. 103031–103063, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does clip's generalization performance mainly stem from high train-test similarity? In *ICLR*, 2024.

Prasanna Mayilvahanan, Roland Zimmermann, Thaddäus Wiedemer, Evgenia Rusak, Attila Juhos, Matthias Bethge, and Wieland Brendel. In search of forgotten domain generalization. In *ICLR*, 2025.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pp. 4990–4999, 2017.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, pp. 102–118, 2016.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, pp. 10745–10755, 2021.

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.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *CVPR*, pp. 14824–14834, 2025.

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

Yihao Xue, Siddharth Joshi, Dang Nguyen, and Baharan Mirzasoleiman. Understanding the robustness of multi-modal contrastive learning to distribution shift. In *ICLR*, 2024.

Juan Yeo, Soonwoo Cha, Jiwoo Song, Hyunbin Jin, and Taesup Kim. ATAS: Any-to-any self-distillation for enhanced open-vocabulary dense prediction. In *ICCV*, pp. 20390–20400, 2025.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pp. 2633–2642, 2020.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. In *NeurIPS*, pp. 32215–32234, 2023.

Xi Yu, Shinjae Yoo, and Yuewei Lin. CLIPCEIL: Domain generalization through CLIP via channel refinement and image-text alignment. In *NeurIPS*, 2024.

Yitian Zhang, Xu Ma, Yue Bai, Huan Wang, and Yun Fu. Accessing vision foundation models via imagenet-1k. In *ICLR*, 2025.

Dong Zhao, Qi Zang, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Open-vocabulary domain generalization in urban-scene segmentation. In *CVPR*, pp. 20388–20398, 2026.

Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. In *NeurIPS*, pp. 338–350, 2022.

APPENDIX

This appendix provides the following details.

A Implementation Details **13**

A.1 Model Architecture and Hyperparameters 13

A.2 Training Details 14

A.3 Inference Steps 14

B Technical Details of Sec. 2.1 **14**

B.1 Representation Similarity Based on CKA 14

B.2 Computing Shared-Neuron Similarity 15

C Additional Experiments **16**

C.1 Effect of Training-Region Construction 16

C.2 Performance of Various VFMs 16

C.3 Impact of Hyperparameters 16

C.4 Analysis of Fine-Tuned Decoders 17

C.5 Per-Class Performance 17

D Computational Cost **17**

E Qualitative Results **18**

F Limitations **21**

A IMPLEMENTATION DETAILS

A.1 MODEL ARCHITECTURE AND HYPERPARAMETERS

We use EVA02-CLIP-B-16 (Sun et al., 2023) as the student image encoder and a frozen DINOv2-reg (Oquab et al., 2023; Darcet et al., 2024) as the SSL-pretrained VFM. The 12-block CLIP encoder is routed at blocks $\mathcal{S} = \{4, 8, 12\}$. At each selected block l, the teacher and student routers are independent two-layer MLPs. They map the VFM consensus and pooled CLIP patch-token features, respectively, to the C_l expansion channels of that block's MLP: $D_f \to 256 \to C_l$ for $\mathcal{R}_T^l$ and $D_l \to 256 \to C_l$ for $\mathcal{R}_S^l$. Each router applies LayerNorm and GELU after its first linear layer, followed by a linear output layer and a row-wise softmax over the C_l channels. The VFM remains frozen; the CLIP image encoder and both sets of routers are trainable.

The resulting gates modulate the normalized hidden activations of the selected CLIP MLPs before their output projections, leaving the attention layers unchanged. We set the gate strength in Eq. 6 to $\alpha = 0.5$ and the cross-view routing weight in Eq. 5 to $\beta = 1$. For masked reconstruction, we randomly mask 40% of the CLIP patches. A trainable projection maps the masked, teacher-routed CLIP dense features into the VFM feature space before region pooling. We set $\lambda_{\text{route}} = \lambda_{\text{swap}} = \lambda_{\text{recon}} = 1$ in Eq. 9.

A.2 Training Details

We train ReoNa on the COCO train2017 split (Lin et al., 2014), using its source annotations to construct the GT regions. Images are processed at 560×560 pixels. We use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 1×10^{-5} and a weight decay of 0.1. Training runs for 12 epochs with a total batch size of 16 on two NVIDIA A800 GPUs. For the decoder-based evaluations, we follow S^2-Corr (Zhao et al., 2026) to train the decoder on GTAV-7 or Cityscapes-7 for the Synthetic-to-Real and Real-to-Real settings, respectively.

A.3 Inference Steps

Inference uses the trained CLIP image encoder and student routers, without the VFM, teacher routers. It consists of an unrouted pass for region discovery and routing prediction, followed by a routed pass for dense feature extraction.

Soft region construction. For each test image window x_i, the first pass produces a dense feature map of shape $512 \times h \times w$ and intermediate patch-token representations $U_i^l \in \mathbb{R}^{N_l \times 768}$ at blocks $l \in \mathcal{S} = \{4, 8, 12\}$. We compute similarities between the dense features and the candidate text embeddings for $\mathcal{C}$, then apply a softmax over classes at each spatial location. Classes are ranked by their mean probability over the window, and the top $R_i = \min(20, |\mathcal{C}|)$ are retained. Their probability maps form the soft region memberships $M_i \in [0,1]^{R_i \times h \times w}$, corresponding to a batched tensor of shape $B \times R_i \times h \times w$ for batch size B.

Visual routing and gate construction. At each selected block, we resize the soft memberships to its token grid and flatten them as $M_i^l \in [0,1]^{R_i \times N_l}$. Membership-weighted pooling gives the region features used by the student router:

$$A_{i,r}^l = \frac{\sum_{n=1}^{N_l} M_{i,r,n}^l U_{i,n}^l}{\sum_{n=1}^{N_l} M_{i,r,n}^l + \epsilon}, \qquad Q_i^{S,l} = \mathrm{softmax}_{C_l}\left(\mathcal{R}_S^l(A_i^l)\right), \tag{10}$$

where $A_i^l \in \mathbb{R}^{R_i \times 768}$ and $Q_i^{S,l} \in \mathbb{R}^{R_i \times C_l}$, with $C_l = 2048$. We normalize each token's memberships across regions to obtain $\bar{M}_i^l \in [0,1]^{N_l \times R_i}$ and construct $G_i^l = \mathcal{G}_i^l(Q_i^{S,l})$ using Eq. 6. Each patch gate has 2048 channel weights; uniform routing yields an all-ones gate, and the CLS-token gate is fixed to one.

Routed dense prediction. The second pass reuses the first-pass tokens immediately before the earliest routed block and recomputes the remaining blocks. At each selected block, G_i^l multiplies the normalized MLP hidden activations before their output projection. This produces routed dense features of shape $512 \times h \times w$ per window, which are used for matching against the full vocabulary $\mathcal{C}$ in zero-shot evaluation or as inputs to the finetuned segmentation decoder.

B Technical Details of Sec. 2.1

B.1 Representation Similarity Based on CKA

We use linear centered kernel alignment (CKA) (Kornblith et al., 2019) to compare dense representations between domains at each visual layer. For each encoder separately, we compute one CKA score per domain pair and layer from sampled patch features concatenated across images.

Feature extraction. We evaluate `EVA02-CLIP-B-16` and `dinov2_vitb14_reg` at an input size of 560. From each Transformer block, we retain the spatial patch tokens, excluding the class token from EVA-CLIP and the class and register tokens from DINOv2-reg. For the final dense representation, we use the EVA-CLIP dense encoder output and the DINOv2-reg patch tokens before the final LayerNorm.

Sampling and feature matrices. We sample up to 64 images independently from each domain. For an image with N spatial tokens, we retain $P = 128$ evenly spaced indices from the token sequence:

$$j_k = \left\lfloor \frac{k(N-1)}{P-1} \right\rfloor, \qquad k = 0, \ldots, P-1, \tag{11}$$

Features are concatenated in image and sampled-token order, keeping at most the first $M = 8192$ rows per domain and layer. This yields $X_a^l \in \mathbb{R}^{m_a \times d_l}$ and $X_b^l \in \mathbb{R}^{m_b \times d_l}$ for domains a and b, where d_l is the feature dimension. Before comparison, both matrices are truncated to their first $n = \min(m_a, m_b)$ rows, giving $X, Y \in \mathbb{R}^{n \times d_l}$.

Centering and linear CKA. Each channel is centered over all retained token rows, including rows from different images:

$$\widetilde{X} = H_n X, \qquad \widetilde{Y} = H_n Y, \qquad H_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top, \tag{12}$$

where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix and $\mathbf{1}_n \in \mathbb{R}^n$ is the all-ones vector.

We compute linear CKA in feature space:

$$\mathrm{CKA}(X, Y) = \frac{\|\widetilde{X}^\top \widetilde{Y}\|_F^2}{\|\widetilde{X}^\top \widetilde{X}\|_F \|\widetilde{Y}^\top \widetilde{Y}\|_F}, \tag{13}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Higher CKA values indicate more similar representational structure between domains.

Aggregation and interpretation. For each layer, CKA is computed once over the retained feature matrices for every unordered pair of domains. The weather curves in Fig. 3 compare each ACDC adverse-weather domain with its associated normal-condition reference domain (Sakaridis et al., 2021).

B.2 Computing Shared-Neuron Similarity

We measure shared-neuron similarity by comparing the most influential feature channels across domains. The procedure is applied separately to each encoder and covers all its L Transformer blocks. A neuron is identified by its block index l and channel index c in the intermediate patch-token representations $U^l(x) \in \mathbb{R}^{N_l \times D_l}$. Spatial tokens and final dense features are extracted as described in Appendix B.1.

Neuron importance. For each domain a, we use up to 16 batches of two sampled images, denoted by $\{\mathcal{B}_{a,k}\}_{k=1}^{K_a}$. Let $z_p(x) \in \mathbb{R}^{D_{\text{out}}}$ be the final dense feature at spatial position p, with N_{out} positions. We average these features spatially and define the attribution objective as their mean squared magnitude:

$$\bar{z}(x) = \frac{1}{N_{\text{out}}} \sum_{p=1}^{N_{\text{out}}} z_p(x), \qquad e(\mathcal{B}) = \frac{1}{|\mathcal{B}| D_{\text{out}}} \sum_{x \in \mathcal{B}} \|\bar{z}(x)\|_2^2. \tag{14}$$

The importance of channel c at block l is the absolute activation–gradient product, averaged over images and spatial tokens within each batch and then over batches:

$$s_{a,c}^l = \frac{1}{K_a} \sum_{k=1}^{K_a} \mathrm{mean}_{x \in \mathcal{B}_{a,k},\, p} \left| U_{p,c}^l(x) \frac{\partial e(\mathcal{B}_{a,k})}{\partial U_{p,c}^l(x)} \right|. \tag{15}$$

This produces a domain-specific importance vector $s_a^l \in \mathbb{R}^{D_l}$ for each block.

Neuron selection and Jaccard overlap. Within each domain and block, we retain the $k_l = \lceil 0.1 D_l \rceil$ channels with the largest importance scores. Their indices form the selected neuron set $\mathcal{I}_a^l \subseteq \{1, \ldots, D_l\}$. For domains a and b, we compare the selected sets at the same block of the same encoder using

$$J_l(a, b) = \frac{|\mathcal{I}_a^l \cap \mathcal{I}_b^l|}{|\mathcal{I}_a^l \cup \mathcal{I}_b^l|}. \tag{16}$$

Higher overlap indicates that more of the same influential channels are shared across the two domains.

Average neuron similarity. We take the equally weighted mean of the Jaccard scores over all L Transformer blocks:

$$\overline{J}(a, b) = \frac{1}{L} \sum_{l=1}^{L} J_l(a, b). \tag{17}$$

Table 5: Effect of training-region construction on ReoNa. Results are reported in mIoU.

	Dv-19					Dv-58				
Method	CS-19	ACDC-19	BDD-19	Mapi-19	Avg.	ACDC-41	BDD-41	Mapi-30	RW-10	Avg.
SAM Mask Generator	44.0	40.7	42.7	46.9	43.6	41.9	48.1	37.1	28.5	38.9
Region Proposals	42.6	41.2	42.7	46.8	43.3	44.3	48.6	37.2	28.9	39.8
Cluster on VFM Features	42.6	38.9	42.8	47.0	42.8	42.0	48.9	37.0	28.3	39.1
GT	44.2	40.7	43.8	47.2	44.0	42.4	48.8	37.5	28.8	39.4

Table 6: Effect of the VFM teacher on ReoNa. Results are reported in mIoU.

	Dv-19					Dv-58				
VFM teacher	CS-19	ACDC-19	BDD-19	Mapi-19	Avg.	ACDC-41	BDD-41	Mapi-30	RW-10	Avg.
DINOv2-B	40.3	34.9	37.7	41.2	38.5	39.4	46.3	33.0	25.6	36.1
DINOv2-reg-B	44.2	40.7	43.8	47.2	44.0	42.4	48.8	37.5	28.8	39.4
DINOv3-B	44.7	40.1	43.4	47.1	43.8	42.0	49.4	36.7	27.9	39.0

Each entry in the shared-neuron similarity matrix is $\overline{J}(a, b)$ for one domain pair. The matrix is symmetric with unit diagonal, and its entries lie in $[0, 1]$. Figure 4 reports these averages for dataset pairs and ACDC adverse-weather/reference pairs.

C Additional Experiments

C.1 Effect of Training-Region Construction

We examine how the choice of training regions affects ReoNa, comparing four sources: automatic masks from SAM (Kirillov et al., 2023), region proposals, clusters of VFM features, and ground-truth (GT) regions. Following CLIPSelf (Wu et al., 2024), we obtain region proposals from a region proposal network (RPN) trained on COCO's train2017 split. For feature clustering, we apply DBSCAN (Ester et al., 1996) to the DINOv2-reg dense features of each image. GT regions are derived directly from the source annotations y_i, as in our default configuration. In each variant, the resulting regions define the masks M_i used for consensus pooling and region-conditioned routing. The same masks are shared by the spatially aligned views x_i and $\hat{x}_i$.

Table 5 shows comparable performance across the four choices, with average results ranging from 42.8 to 44.0 mIoU on Dv-19 and from 38.9 to 39.8 mIoU on Dv-58. GT regions yield the highest average on Dv-19 and reach 39.4 mIoU on Dv-58, within 0.4 mIoU of the best-performing region proposals. These results suggest that ReoNa is not strongly dependent on a particular region-generation method. We adopt GT regions by default because they provide competitive performance while avoiding additional mask generation or feature clustering during training.

C.2 Performance of Various VFMs

We evaluate the choice of VFM teacher by training ReoNa with DINOv2-B (Oquab et al., 2023), DINOv2-reg-B (Darcet et al., 2024), and DINOv3-B (Siméoni et al., 2025). Table 6 reports the resulting segmentation performance of ReoNa. Replacing DINOv2-B with DINOv2-reg-B increases the average performance from 38.5 to 44.0 mIoU on Dv-19 and from 36.1 to 39.4 mIoU on Dv-58. DINOv3-B yields comparable averages of 43.8 and 39.0 mIoU, respectively, and achieves the highest results on CS-19 and BDD-41 among the three variants. DINOv2-reg-B retains the highest average on both benchmarks and is therefore used as our default VFM, while DINOv3-B provides a similarly effective alternative.

C.3 Impact of Hyperparameters

Gate strength α. Figure 8 examines the effect of the gate strength in Eq. 6. The performance remains stable across the tested values on both benchmarks, showing that ReoNa is robust to the choice of α.

Table 7: Synthetic-to-Real performance with the CAT-Seg decoder and ViT-B/16 image encoders. Results are reported in mIoU.

	Dv-19					Dv-58				
Method	CS-19	ACDC-19	BDD-19	Mapi-19	Avg.	ACDC-41	BDD-41	Mapi-30	RW-10	Avg.
CAT-Seg (Cho et al., 2024)	43.7	37.6	45.4	48.8	43.9	51.2	56.2	39.3	35.7	45.6
DeCLIP+CAT-Seg (Wang et al., 2025)	45.5	37.9	46.8	50.9	45.3	52.9	58.0	43.5	36.3	47.7
ReoNa+CAT-Seg	47.2	42.0	49.0	55.4	48.4	55.1	60.4	44.6	38.8	49.7

Mask ratio. Figure 9 varies the fraction of randomly masked patches in the reconstruction objective from 0.2 to 0.6. Both benchmarks achieve their best results at a mask ratio of 0.4, with 44.0 mIoU on Dv-19 and 39.4 mIoU on Dv-58.

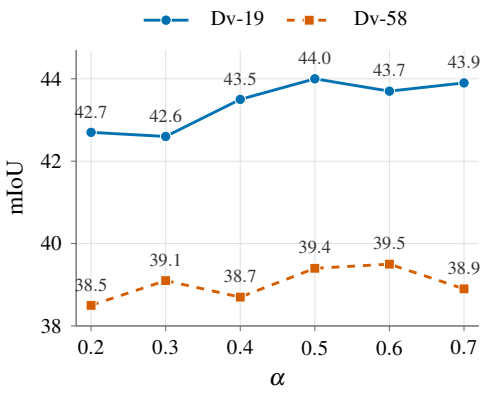

Figure 8: Effect of gate strength α on average mIoU.

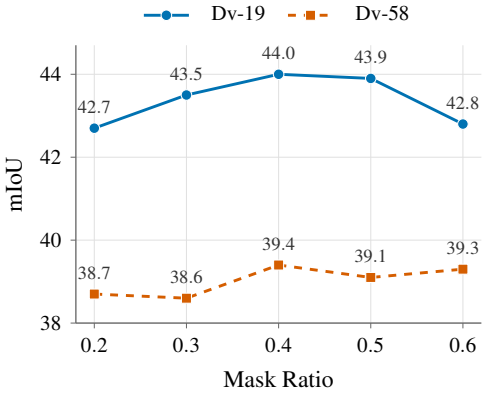

Figure 9: Effect of mask ratio on average mIoU.

C.4 Analysis of Fine-Tuned Decoders

S^2-Corr (Zhao et al., 2026) is specifically optimized for domain-generalized semantic segmentation. To isolate gains attributable to this decoder design, we replace it with CAT-Seg (Cho et al., 2024), which is developed for general open-vocabulary semantic segmentation. We fine-tune the CAT-Seg decoder on GTAV-7 under the Synthetic-to-Real setting and evaluate its combinations with DeCLIP and ReoNa. As shown in Table 7, ReoNa+CAT-Seg achieves average results of 48.4 mIoU on Dv-19 and 49.7 mIoU on Dv-58, exceeding DeCLIP+CAT-Seg by 3.1 and 2.0 mIoU, respectively. The improvements across all test sets show that the benefits of ReoNa's visual representations extend to a general OVSS decoder and do not rely on the DG-specific design of S^2-Corr.

C.5 Per-Class Performance

Dv-19. Table 8 reports the per-class IoU of ReoNa on CS-19, ACDC-19, BDD-19, and Mapi-19 using their shared urban-scene vocabulary.

Dv-58. Table 9 presents the per-class results on ACDC-41 and BDD-41, covering urban-scene and additional object categories. Tables 10 and 11 report the results for the dataset-specific categories of Mapi-30 and RW-10, respectively.

D Computational Cost

ReoNa introduces additional routing modules to several layers of the CLIP image encoder, increasing the parameter count from 80.0M for DeCLIP to 87.0M, an overhead of 8.8%. At the reported precision, ReoNa requires 1.5T FLOPs for the forward pass and 3.0T FLOPs for the backward pass.

Table 8: Per-class IoU of ReoNa on Dv-19.

Class	CS-19	ACDC-19	BDD-19	Mapi-19
road	88.04	78.72	88.05	87.49
sidewalk	43.97	28.87	43.79	54.55
building	73.80	75.46	71.74	72.30
wall	11.53	26.53	15.21	22.94
fence	38.97	31.20	38.75	45.36
pole	10.82	12.35	9.81	11.21
traffic light	33.22	26.48	24.00	33.69
traffic sign	30.67	35.64	29.09	46.29
vegetation	75.06	42.68	56.56	59.32
terrain	0.00	1.13	1.50	0.45
sky	77.64	77.87	78.93	87.38
person	58.67	42.78	53.88	53.46
rider	14.01	17.01	16.09	15.61
car	74.87	63.71	76.21	68.41
truck	37.90	41.61	50.36	38.37
bus	54.44	45.71	71.25	70.15
train	11.83	51.89	0.00	26.16
motorcycle	45.29	32.52	52.30	48.84
bicycle	58.81	40.98	54.22	54.24

Table 9: Per-class IoU of ReoNa on ACDC-41 and BDD-41 in Dv-58.

Urban-scene classes			Additional classes		
Class	ACDC-41	BDD-41	Class	ACDC-41	BDD-41
road	75.88	85.31	bag	42.06	62.85
sidewalk	22.82	37.49	ball	56.73	60.98
building	74.91	70.62	barrel	37.28	47.88
wall	22.68	19.47	bench	30.11	56.35
fence	25.61	27.02	bird	54.81	68.22
pole	12.02	9.35	bottle	75.83	73.83
traffic light	18.89	26.16	cart	40.99	41.15
traffic sign	25.30	25.41	cat	70.75	56.67
vegetation	38.56	55.39	chair	50.76	57.28
terrain	0.45	3.25	cow	63.56	68.24
sky	74.04	79.31	deer	46.00	55.67
person	22.71	26.34	dog	58.04	66.22
rider	12.23	16.35	drone	0.45	0.00
car	55.44	63.27	elephant	69.92	69.37
truck	47.90	45.99	hat	60.27	72.70
bus	34.06	76.11	horse	34.81	67.61
train	42.04	0.00	robot	34.27	31.76
motorcycle	23.99	57.08	sheep	55.08	77.81
bicycle	25.60	35.37	table	22.30	40.26
			toy	51.18	33.59
			umbrella	68.53	79.48
			zebra	59.52	54.98

The VFM and teacher routers are used only during training and are removed at inference, leaving the CLIP image encoder and student routers for deployment.

E Qualitative Results

Segmentation map. As shown in Fig. 10, ReoNa produces cleaner and more spatially coherent predictions than DeCLIP across diverse urban scenes. In particular, it reduces fragmented predictions

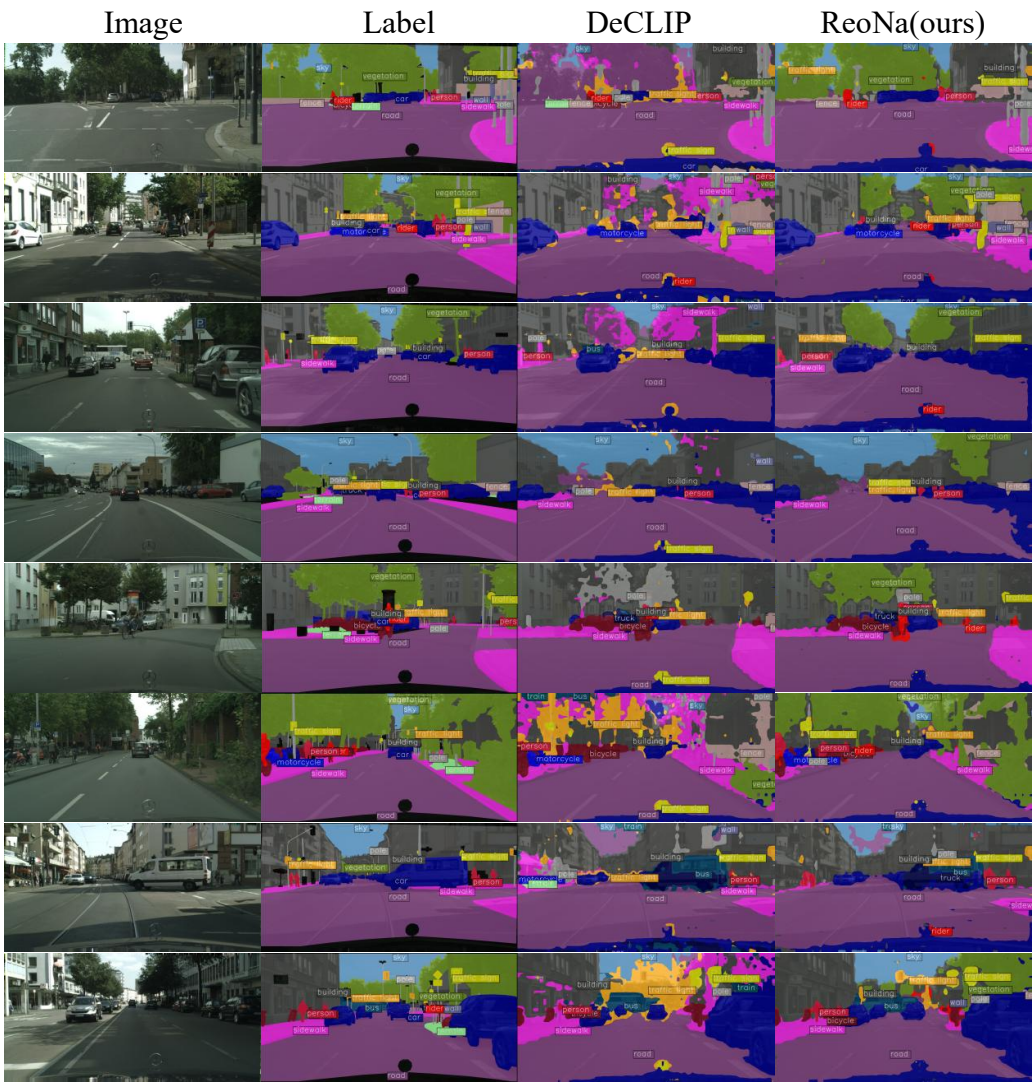

Figure 10: Qualitative comparison on unseen urban scenes. From left to right: input image, ground-truth label, DeCLIP prediction, and ReoNa prediction.

Table 10: Per-class IoU of ReoNa on Mapi-30.

Class	IoU	Class	IoU
road	75.18	sand	17.37
sidewalk	52.09	person	51.32
building	73.01	rider	17.89
wall	21.22	car	65.14
bridge	32.55	truck	33.72
tunnel	9.05	bus	55.17
traffic sign	8.31	train	31.21
traffic light	31.56	bicycle	49.74
pole	10.81	motorcycle	48.49
fence	38.96	animal	44.97
sky	85.02	signboard	25.63
vegetation	58.06	railway	11.89
terrain	4.06	boat	21.98
water	55.13	chair	23.76
snow	57.14	trash can	14.39

Table 11: Per-class IoU of ReoNa on RW-10.

Class	IoU
road	86.38
sidewalk	48.55
barrier	54.66
police vehicle	2.41
work vehicle	47.43
police officer	1.84
worker	38.20
cone	5.59
arrow board	0.00
ttc sign	2.45

within large regions such as roads, buildings, and vegetation, while preserving clearer boundaries and more complete predictions for vehicles, pedestrians, and riders. These examples qualitatively demonstrate the improved robustness of ReoNa's dense predictions under domain shifts.

Layer-wise intermediate feature consistency. Figures 11 and 12 compare how the intermediate patch-token embeddings evolve for the same scene under an appearance shift. DeCLIP exhibits visibly separated, domain-specific clusters at several intermediate and later blocks, indicating that its internal representations remain sensitive to the appearance change. In contrast, the original and shifted embeddings of ReoNa are more strongly intermingled and preserve more similar geometric structure across most layers, including the final blocks. Although UMAP provides a qualitative view, this layer-wise pattern is consistent with improved cross-domain sharing throughout the CLIP image encoder rather than alignment confined to its final dense features.

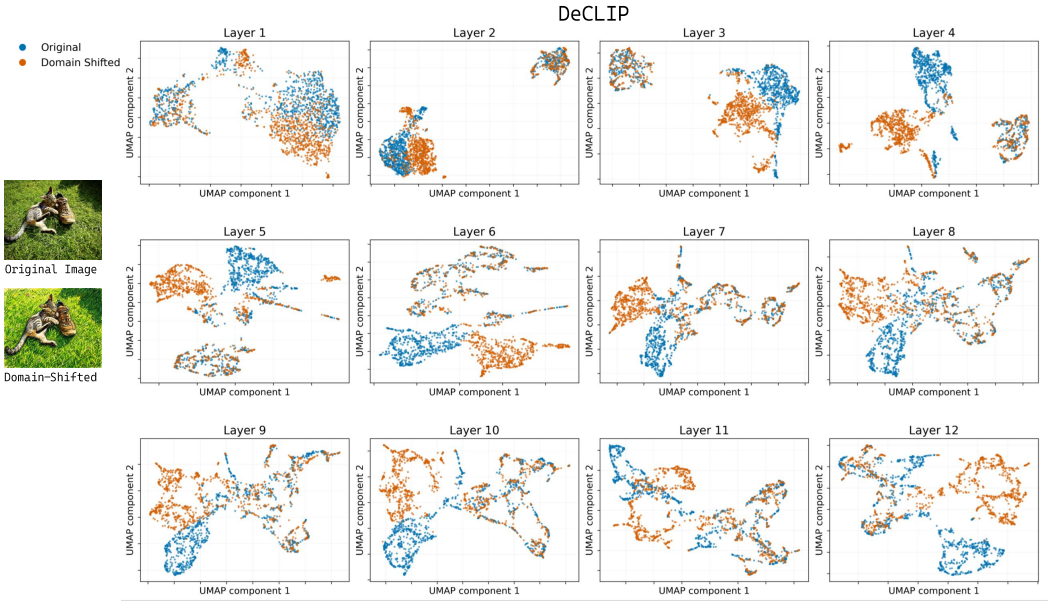

Figure 11: **Layer-wise intermediate embeddings of DeCLIP.** UMAP projections of patch-token features from the original image (blue) and its domain-shifted counterpart (orange) across the 12 visual blocks.

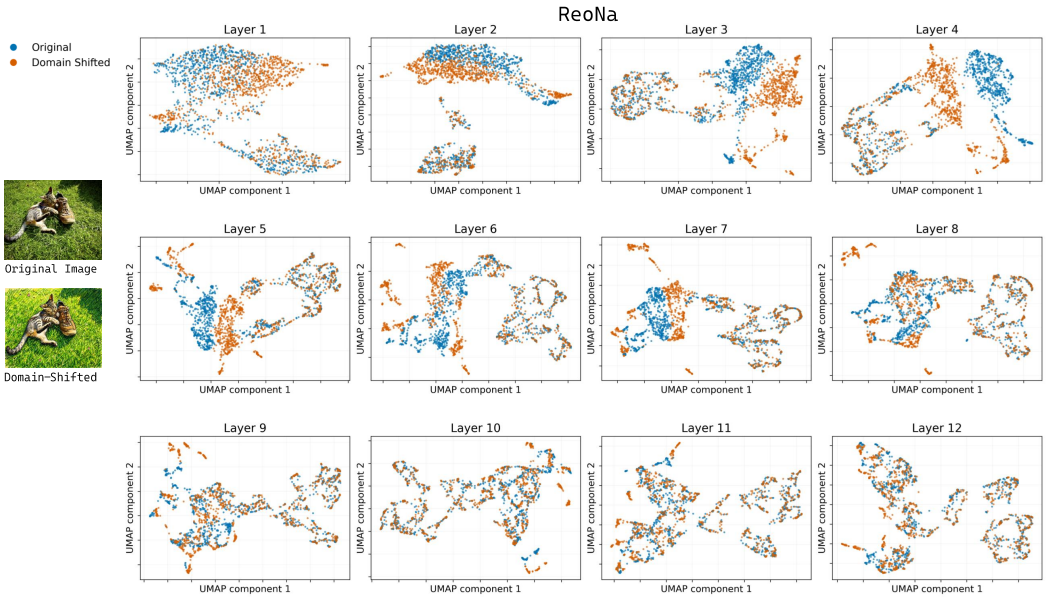

Figure 12: **Layer-wise intermediate embeddings of ReoNa.** UMAP projections of patch-token features from the original image (blue) and its domain-shifted counterpart (orange) across the 12 visual blocks.

F LIMITATIONS

ReoNa constructs cross-view consensus from appearance transformations that preserve spatial layout, allowing the same regions to be associated reliably across views. This design isolates appearance-induced domain variation and avoids introducing an additional region-matching module during training. A natural extension is to incorporate geometry-changing transformations by replacing coordinate-shared regions with explicitly matched region correspondences, while retaining the same consensus and routing objectives.