DIVEUP: LEARNING FEATURE UPSAMPLING FROM DIVERSE VISION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature upsampling has emerged as an effective way to adapt the vision foundation models (VFMs) to pixel-level understanding tasks. However, existing backbone-agnostic upsampling approaches are typically trained through single-model self-reconstruction, making them susceptible to model-specific spatial misalignment and feature artifacts. Addressing this, we introduce DiveUp, a novel unified multi-VFM feature upsampling framework that jointly learns a single shared upsampler on multiple diverse VFMs, improving its generalization across heterogeneous and unseen VFMs. At each training step, DiveUp samples one VFM, reconstructs its native high-resolution features with the shared upsampler, and leverages the corresponding spatially aligned teacher to provide cross-model geometric guidance. To provide additional spatial supervision beyond self-reconstruction, we introduce a local center-of-mass (COM) field derived from local feature affinities as a relational representation of geometric structure. The COM field transfers teacher geometry without directly aligning heterogeneous VFM feature spaces, improving the quality of upsampled features. To further understand how the composition of training VFMs affects generalization, we conduct systematic analysis involving 14 diverse VFMs and find that broader representational diversity in the training mixture leads to stronger transfer to unseen models. We further show that explicit geometric guidance is particularly beneficial for backbones with noisy or poorly aligned spatial representations. Experiments across dense prediction tasks demonstrate that DiveUp achieves state-of-the-art zero-shot transfer across diverse VFMs while remaining competitive with VFM-specific upsamplers.

1 INTRODUCTION

Features from pretrained vision foundation models (VFMs) (Tschannen et al., 2025; Siméoni et al., 2025; Oquab et al., 2023; Ranzinger et al., 2024) have been widely applied in vision tasks (El Banani et al., 2024; Huang et al., 2024; Kirillov et al., 2023; Fan et al., 2025; Ravi et al., 2024; Tumanyan et al., 2024) owing to the rich semantics and excellent generality. Nevertheless, due to patch-based tokenization or pooling operation in these pretrained models, the resulting feature maps are usually much smaller (*e.g.*, 16 times smaller) in spatial resolution than the original input images. This significant reduction in resolution leads to the loss of important spatial details (*e.g.*, geometric structures and boundaries) in feature maps, thereby limiting their utility for downstream tasks, such as semantic segmentation and depth estimation, that demand pixel-level understanding for dense prediction.

To mitigate this limitation, recent works (Couairon et al., 2025; Fu et al., 2024; Wimmer et al., 2026; Huang et al., 2025; Chambon et al., 2025) learn to upsample VFM features, typically through self-reconstruction using features from a single source VFM. While this preserves the native semantics of the source model, it can cause the learned upsampler to overfit to source-specific feature characteristics and spatial artifacts. Consequently, model-specific deficiencies such as spatial misalignment (Wang et al., 2023) and high-norm artifacts (Darcet et al., 2024; Yang et al., 2024) may be inherited by the upsampled features, degrading its performance. This issue is especially pronounced for multimodal VFMs (Tschannen et al., 2025; Radford et al., 2021), whose features may exhibit weaker spatial localization due to vision-language alignment. Their high-norm patch artifacts survive upsampling: existing upsamplers propagate them into uninformative regions and fragment thin structures such as an aircraft's wings or a motorcycle's outline (Fig. 3).

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.

Thus motivated, we propose DiveUp, which contains two mechanisms to correct inherited spatial misalignments: (1) diversifying the training sources to prevent semantic overfitting, and (2) enforcing explicit structural guidance to preserve geometric details. We found that, while each mechanism independently improves overall representation quality, their combination is essential for restoring accurate semantic boundaries. At each training step, DiveUp samples one VFM to provide native high-resolution targets, while a fixed, spatially aligned teacher provides cross-model geometric guidance. By dynamically resampling diverse VFMs to jointly optimize a single shared upsampler, DiveUp avoids overfitting to any individual source model's artifacts, thereby encouraging robust zero-shot generalization across heterogeneous architectures. Furthermore, because standard self-reconstruction preserves the sampled VFM's native semantics but supervises spatial structure only implicitly, we explicitly decouple the spatial alignment via the teacher-guided geometric branch. Its signal is a local center-of-mass (COM) field, defined as the first moment of a local self-affinity distribution. As a universal relational quantity, the COM field is inherently comparable across models of varying feature widths, ensuring that DiveUp transfers the teacher's local *arrangement* without directly aligning heterogeneous feature spaces.

To better understand the factors affecting generalization, we conduct systematic analyses involving 14 diverse VFMs and uncover two key findings. First, under a fixed backbone budget, training mixtures with greater representational diversity, quantified by their centered kernel alignment (CKA) span, consistently show stronger transfer to unseen architectures. Second, explicit spatial guidance provides larger benefits for backbones with noisier or less well-aligned spatial representations, such as CLIP-based models and CNNs. These findings provide practical insights into both the composition of training VFMs and the use of geometric guidance for robust feature upsampling. To validate the effectiveness of DiveUp, we conduct experiments across diverse VFMs, architectures, and dense prediction tasks. The results show that DiveUp achieves state-of-the-art results among VFM-agnostic upsamplers while maintaining strong zero-shot generalization to unseen architectures, without requiring per-model retraining.

In summary, our contributions are as follows: (1) We propose DiveUp, an encoder-agnostic framework that jointly learns a shared upsampler from diverse VFMs; (2) We introduce the local center-of-mass (COM) field, a relational quantity that is comparable across models of different feature widths, and use it to transfer a teacher's local spatial *arrangement* without transferring its representation; (3) A systematic study over 14 VFMs shows that the range of pretraining paradigms a mixture covers is a stronger predictor of zero-shot transfer than its size, and that spatial guidance benefits structurally noisy backbones most; (4) Extensive experiments demonstrate state-of-the-art performance among VFM-agnostic upsamplers and strong zero-shot generalization to unseen architectures.

2 Related Work

VFM-Specific Feature Upsampling. Feature upsampling has attracted increasing attention in recent years. FeatUp (Fu et al., 2024) and FeatSharp (Ranzinger et al., 2025) recover spatial details through multi-view consistency, using a trained Joint Bilateral Upsampler or test-time implicit representation in FeatUp and patch tiling with local transformer attention in FeatSharp. Opting for a more straightforward and efficient design, LiFT (Suri et al., 2024) adopts a fully convolutional U-Net architecture for spatial restoration. Building upon this to overcome the issue of semantic drift during iterative upsampling, the recent UPLiFT (Walmer et al., 2026) proposes a novel Local Attender operator. By computing attention over fixed directional offsets, UPLiFT ensures feature consistency with linear computational complexity. Taking a different route, LoftUp (Huang et al., 2025) employs a coordinate-based cross-attention transformer trained in two stages via SAM-generated (Kirillov et al., 2023) pseudo-labels. Alongside these, JAFAR (Couairon et al., 2025) adopts a lightweight cross-attention module that optimizes feature reconstruction at low upsampling ratios to generalize to arbitrary scales. However, the defining bottleneck of these approaches is that they are inherently *VFM-specific*. Their reliance on fixed, model-tied feature spaces for every individual foundation model, fundamentally restricting them from generalizing zero-shot to diverse, unseen VFMs.

VFM-Agnostic Feature Upsampling. To overcome the retraining bottleneck, recent works have studied universal, feature-agnostic upsampling. AnyUp (Wimmer et al., 2026) presents a feature-agnostic channel projection to decouple key generation from the source VFM, while NAF (Chambon et al., 2025) further removes the spatial feature transform and relies on image-based guidance for

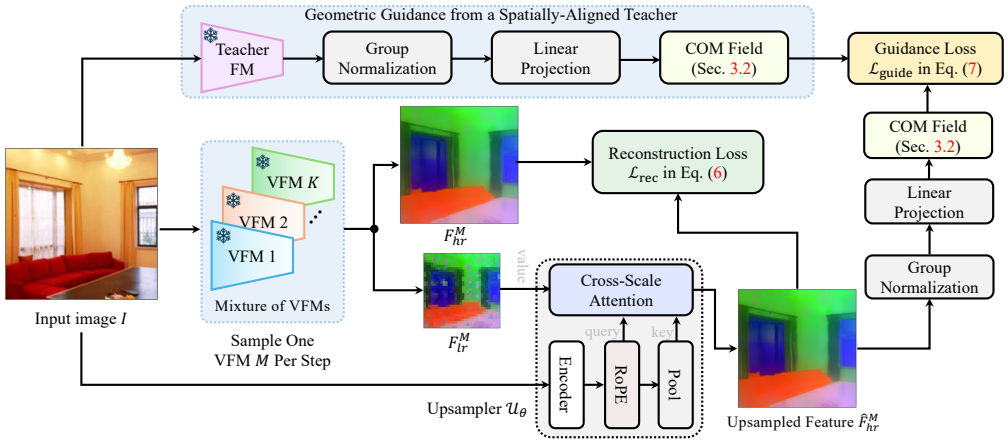


Figure 1: Overview of the DiveUp, which leverages diverse VFMs to learn feature upsampling. *Best viewed in color and by zooming in for all the figures in this paper.*

cross-scale upsampling. Upsample Anything (Seo et al., 2025) instead performs test-time adaptation using a per-image anisotropic Gaussian kernel. Despite their improved generality, these methods still rely on single-model feature reconstruction or source features and may therefore inherit model-specific spatial artifacts, particularly for VFMs with poor dense localization. Our DiveUp also belongs to the VFM-Agnostic category. However, ***different from*** the aforementioned methods that rely on single-model self-reconstruction and thereby overfit to the source model's inherent spatial flaws, DiveUp fundamentally breaks this dependency. By training on a *diverse mixture* of VFMs, our model acquires a robust, backbone-agnostic alignment strategy. Furthermore, rather than blindly inheriting spatial errors, DiveUp leverages a local center-of-mass (COM) field from a *single spatially-aligned teacher*. This provides explicit geometric guidance, effectively decoupling spatial alignment from semantic reconstruction and yielding more robust zero-shot upsampled features.

Classical Handcrafted Upsampling. Early non-learning approaches rely on parameter-free interpolation and edge-aware filtering. Simple methods like bilinear interpolation often ignore image content and thus blur semantic boundaries. To address this, classical filters such as Joint Bilateral Upsampling (JBU) (Kopf et al., 2007) and Guided Image Filtering (GIF) (He et al., 2013) leverage high-resolution RGB images as spatial priors to guide the upsampling of low-resolution signals. While highly efficient and requiring no training, these handcrafted kernels operate fundamentally on low-dimensional spaces (*e.g.*, color or tone) using fixed mathematical forms. Consequently, they struggle to capture the complex, high-dimensional semantic manifolds of modern VFMs.

3 The Proposed Methodology

In this work, we formulate feature upsampling as a multi-VFM relational guidance task rather than standard intra-model reconstruction (Couairon et al., 2025; Wimmer et al., 2026; Huang et al., 2025; Chambon et al., 2025). DiveUp is a universal, plug-and-play training framework that can be integrated with existing upsamplers. It optimizes them through two pathways: semantic reconstruction from a diverse mixture of VFMs and geometric guidance from a local center-of-mass (COM) field extracted from a spatially aligned teacher.

3.1 Problem Formulation and Framework Overview

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a diverse mixture of source vision foundation models $\mathcal{M} = \{M_i\}_{i=1}^{K}$ with K the number of VFMs, we uniformly sample a model $M \in \mathcal{M}$ at each training step. This sampled model extracts a low-resolution feature map $F_{lr}^{M} \in \mathbb{R}^{h\times w\times C_M}$. Our objective is to learn a universal, parameterized upsampler $\mathcal{U}_\theta$ to reconstruct a high-resolution feature map $\hat{F}_{hr}^{M} \in \mathbb{R}^{H_{hr}\times W_{hr}\times C_M}$, in the sampled model's own channel space at an arbitrary target resolution. Specifically, the target spatial dimensions scale as $H_{hr} = s \cdot h$ and $W_{hr} = s \cdot w$ for any given

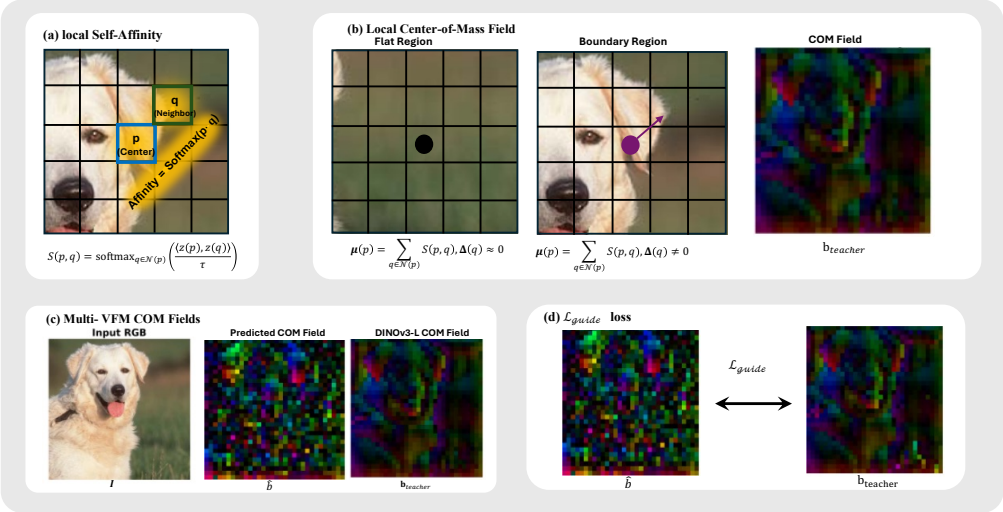


Figure 2: Visualization of the universal relational representation.

upsampling factor s, restoring fine-grained spatial details from coarse patch-level semantics:

$$\hat{F}_{hr}^{M} = \mathcal{U}_\theta(F_{lr}^{M}, I). \tag{1}$$

where $\mathcal{U}_\theta$ computes the feature at position p as an attention-weighted combination of the low-resolution features in its neighborhood $\mathcal{N}(p)$: the raw features F_{lr}^{M} are the values, while queries and keys come from a compact image encoder, so the mixing weights are independent of the source VFM. We instantiate it from NAF (Chambon et al., 2025), but DiveUp is a training framework and any upsampler can take its place. We verify this on two others in Sec. 4. The architecture details are in the **appendix** due to space limitations.

Existing methods reconstruct a single fixed model from itself, so the upsampler inherits that model's spatial misalignments and high-norm artifacts. We break this dependency with two training terms, developed in turn below: a geometric one that aligns the output to a teacher's local center-of-mass field (Sec. 3.2), and a semantic one that reconstructs whichever model is sampled from $\mathcal{M}$ (Sec. 3.3). Neither acts at inference, where $\mathcal{U}_\theta$ carries no teacher and no extra cost; Fig. 1 gives the overview.

3.2 Universal Relational Representation

To explicitly guide the upsampler's spatial alignment without disrupting the native semantic spaces of the diverse training VFMs, we extract a universal relational representation from the structurally robust teacher model (*e.g.*, DINOv3).

Local Self-Affinity. As illustrated in Fig. 2 (a), for each spatial position p in the teacher's feature map, let $\mathcal{N}(p)$ denote a local $w \times w$ window. We first map the raw teacher features $F_{\text{teacher}}(p)$ into a d-dimensional space via a projection head $\phi(\cdot)$. To mitigate the magnitude variations, $\phi(\cdot)$ applies a Group Normalization layer followed by a Linear Projection. This projection yields a compact standardized representation $z(p) = \phi(F_{\text{teacher}}(p))$, reducing the sensitivity of the subsequent similarity computation to feature-magnitude variation. The local structural similarity distribution is defined as follows,

$$S(p,q) = \text{softmax}_{q \in \mathcal{N}(p)}\left(\frac{\langle z(p), z(q)\rangle}{\tau}\right), \quad q \in \mathcal{N}(p), \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the dot product. This distribution captures the teacher's internal geometric understanding, indicating which neighboring pixels share semantic affinity with p.

Local Center-of-Mass (COM) Field. While the affinity matrix captures pairwise relationships, it lacks explicit spatial directionality. As shown in Fig. 2 (b), let $\boldsymbol{\Delta}(q) = (\Delta x_q, \Delta y_q)$ be the relative spatial offset of neighbor q in window coordinates. The center-of-mass (COM) vector at position p is computed as the expected spatial offset via

$$\boldsymbol{\mu}(p) = \sum_{q \in \mathcal{N}(p)} S(p,q)\, \boldsymbol{\Delta}(q). \tag{3}$$

What the COM field measures. On an idealized image with piecewise constant projected features, $\boldsymbol{\mu}(p) = \mathbf{0}$ in region interiors. If the window straddles a boundary and $\langle z_a, z_a \rangle > \langle z_a, z_b \rangle$, it points toward the centroid of the same-region support; for a locally straight boundary, this is the inward normal. Thus the field is boundary-sensitive and silent in interiors. Normalization in $\phi(\cdot)$ mitigates magnitude-driven inversions; precise statements and proofs are in the **appendix**.

In practice, we further normalize the expected spatial offset into a bounded vector field $\mathbf{b}_{\text{teacher}}(p) = \text{clip}(\boldsymbol{\mu}(p)/r, -1, 1)$ with $r = \lfloor w/2 \rfloor$. This forms a universal geometric representation that distills the teacher's boundary awareness independently of its semantic channel dimensions.

3.3 Learning from a Diverse Mixture of VFMs

Existing approaches typically train their upsamplers on a single source model, which inadvertently forces the network to overfit to that model's specific spatial artifacts. DiveUp instead optimizes *one* set of upsampler weights θ against the whole mixture introduced in Sec. 3.1,

$$\mathcal{M} = \{M_1, \ldots, M_K\}, \qquad K = |\mathcal{M}|, \tag{4}$$

resampled at every training step. As demonstrated in our analysis, greater representational diversity in the mixture is strongly associated with improved zero-shot generalization to unseen architectures.

At each step, we uniformly draw $M \sim \text{Unif}(\mathcal{M})$ and employ the shared upsampler to reconstruct that model's high-resolution features F^M_{hr} from its own low-resolution ones via Eq. (1). Simultaneously, we extract the structural COM field $\mathbf{b}_{\text{teacher}}$ from the frozen spatial teacher, which is *shared by all K branches and carries no dependence on M*: geometry is common to every VFM, only semantics are model-specific. To enforce geometric alignment without disrupting the native semantic space of the sampled model, we compute the COM fields in a low-dimensional spatial subspace.

The total objective is therefore an expectation over the mixture, as follows,

$$\mathcal{L}_{\text{total}}(\theta) = \mathbb{E}_{M \sim \text{Unif}(\mathcal{M})} \Big[\underbrace{\mathcal{L}_{\text{rec}}\big(\hat{F}^M_{hr}, F^M_{hr}\big)}_{\text{semantics of } M} + \lambda^M_{\text{guide}} \underbrace{\mathcal{L}_{\text{guide}}\big(\hat{\mathbf{b}}^M, \mathbf{b}_{\text{teacher}}\big)}_{\text{geometry, one shared teacher}} \Big]. \tag{5}$$

The reconstruction loss preserves the native semantic identity of the sampled model. It is applied to ℓ_2-normalized features as follows,

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left\| \frac{\hat{F}^M_{hr}(p)}{\|\hat{F}^M_{hr}(p)\|_2} - \frac{F^M_{hr}(p)}{\|F^M_{hr}(p)\|_2} \right\|^2_2. \tag{6}$$

The relational guidance loss $\mathcal{L}_{\text{guide}}$ applies an L_1 penalty between the student and teacher COM fields:

$$\mathcal{L}_{\text{guide}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \big\| \hat{\mathbf{b}}(p) - \mathbf{b}_{\text{teacher}}(p) \big\|_1, \tag{7}$$

where Ω is the set of all locations and $\hat{\mathbf{b}}^M$ applies the same COM operator to $\hat{F}^M_{hr}$ through a training-only head $\phi^s_{C_M} : \mathbb{R}^{C_M} \to \mathbb{R}^{d_\phi}$, shared by backbones of equal width and discarded at inference (App. S6.2). The teacher has its own frozen projection head: only geometry is matched, and the reconstruction term limits semantic drift.

4 Experiments

We assess the effectiveness of DiveUp across multiple VFMs. Following previous methods (Couairon et al., 2025; Wimmer et al., 2026; Chambon et al., 2025), we conduct linear probing on the upsampled features and assess their performance on different dense prediction tasks, including semantic segmentation and depth estimation.

4.1 Experimental Setup

Implementation Details. DiveUp is implemented using PyTorch (Paszke et al., 2019) on a single NVIDIA RTX A6000 GPU. Following (Chambon et al., 2025), we train the upsampler in DiveUp for

Table 1: Linear probing for Semantic Segmentation (mIoU ↑) on Pascal VOC and Depth Estimation (δ_1 ↑) on NYUv2. All methods here are VFM-agnostic: trained once, then applied zero-shot to any backbone; VFM-specific upsamplers need a training run per backbone and are compared in Tab. 2. The best and second-best results are highlighted in **bold** and underlined.

Methods	Semantic Segmentation (Pascal VOC)					Depth Estimation (NYUv2)				
	ConvNeXt-S (CNN)	MaskCLIP (CLIP)	Swin-S (Window)	DINOv2-S (SSL ViT)	m-**VFM Avg.**	ConvNeXt-S (CNN)	MaskCLIP (CLIP)	Swin-S (Window)	DINOv2-S (SSL ViT)	m-**VFM Avg.**
Bilinear	53.78	68.49	59.94	81.69	73.30	68.53	64.11	67.96	81.79	73.40
VFM-Agnostic Upsamplers (Train once, generalize zero-shot to any VFM)										
AnyUp-S[ICLR'26] (Wimmer et al., 2026)	58.01	73.76	64.37	84.08	76.59	69.34	68.05	68.67	83.37	75.29
AnyUp-M[ICLR'26] (Wimmer et al., 2026)	59.37	75.04	65.66	84.01	77.07	70.25	69.51	70.11	83.33	76.05
NAF[CVPR'26] (Chambon et al., 2025)	55.49	73.32	61.51	84.54	76.20	67.25	66.65	66.64	83.24	74.45
DiveUp (ours)	**63.55**	**76.54**	**67.87**	**84.78**	**78.77**	**72.10**	**70.16**	**70.81**	**84.08**	**76.76**

Table 2: Comparison across downstream tasks and datasets. 'V.A.' marks VFM-agnostic models, which train once; the others need a training run per backbone, LoftUp does release a SigLIP checkpoint, but it targets SigLIP-224 rather than the SigLIP2-B/16-512 evaluated here, so its SigLIP-B row (†) is *not directly comparable* and is reported only for completeness. Best and second best are **bold** and underlined.

Methods	V.A.	Semantic Seg. (mIoU ↑)			Depth
		VOC	COCO-Stuff	ADE20K	NYUv2 (δ_1 ↑)
(a) DINOv2-S backbone					
Bilinear	✓	81.69	59.53	40.23	81.79
FeatUp (JBU)[ICLR'24] (Fu et al., 2024)	✗	83.31	61.80	41.42	81.09
LoftUp[ICCV'25] (Huang et al., 2025)	✗	84.50	62.18	40.87	79.32
JAFAR[NeurIPS'25] (Couairon et al., 2025)	✗	84.36	61.65	41.73	81.96
UPLiFT[CVPR'26] (Walmer et al., 2026)	✗	**85.15**	62.53	**43.02**	83.49
AnyUp-S[ICLR'26] (Wimmer et al., 2026)	✓	84.08	62.10	42.12	83.37
AnyUp-M[ICLR'26] (Wimmer et al., 2026)	✓	84.01	62.00	41.82	83.33
NAF[CVPR'26] (Chambon et al., 2025)	✓	84.54	62.16	42.62	83.24
DiveUp (ours)	✓	84.78	**62.72**	42.71	**84.08**
(b) SigLIP-B backbone					
Bilinear	✓	71.93	51.64	36.26	70.54
FeatUp (JBU)[ICLR'24] (Fu et al., 2024)	✗	71.27	55.02	36.19	67.16
LoftUp[ICCV'25] (Huang et al., 2025)	✗	46.62†	39.39†	20.85†	48.54†
JAFAR[NeurIPS'25] (Couairon et al., 2025)	✗	77.37	58.80	40.02	75.46
UPLiFT[CVPR'26] (Walmer et al., 2026)	✗	79.10	59.58	41.65	**79.25**
AnyUp-S[ICLR'26] (Wimmer et al., 2026)	✓	78.56	58.26	40.65	75.41
AnyUp-M[ICLR'26] (Wimmer et al., 2026)	✓	79.20	59.42	41.02	77.02
NAF[CVPR'26] (Chambon et al., 2025)	✓	77.75	57.27	40.41	73.67
DiveUp (ours)	✓	**81.19**	**60.99**	**42.64**	78.11

25K iterations on the ImageNet (Deng et al., 2009) dataset, using the AdamW optimizer (Loshchilov & Hutter, 2019) with a batch size of 4, an initial learning rate of 2e-4, and a weight decay of 1e-5. The image size used in training is 512^2 and downsized to 256^2 resolution for getting low-resolution feature. The local window size for computing the self-affinity and COM field is set to $w = 7$. To learn a robust VFM-agnostic upsampler, we employ a mixed-training strategy that simultaneously applies four backbones during training: DINOv3-B (Siméoni et al., 2025), SigLIP-B (Tschannen et al., 2025), Franca (Venkataramanan et al., 2025), and PE-Core-S (Bolya et al., 2025). For each backbone, the upsampler is guided by relational representations extracted from a single geometrically aligned teacher VFM: DINOv3-L (Siméoni et al., 2025). In the training objective, the guidance weight λ is empirically set to 0.5. Our code and models will be released.

Linear Probing Protocol and Tasks. To assess DiveUp, following (Couairon et al., 2025; Chambon et al., 2025), we conduct linear probing of upsampled features and evaluate their performance on two representative dense prediction tasks, including semantic segmentation and depth estimation.

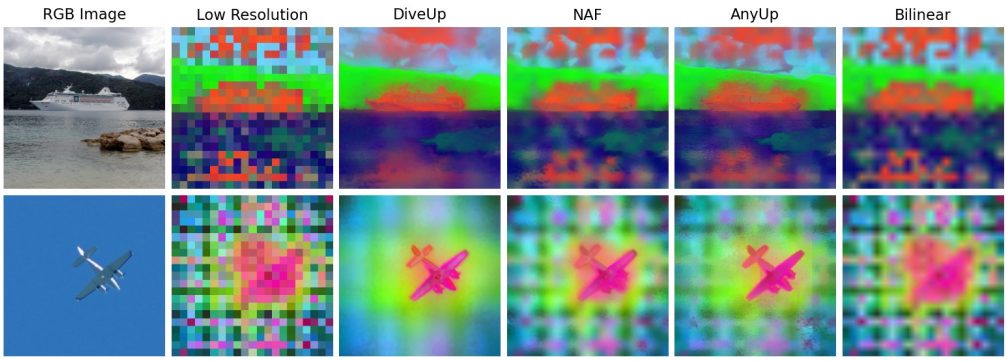


Figure 3: Comparison of upsampled features from different methods.

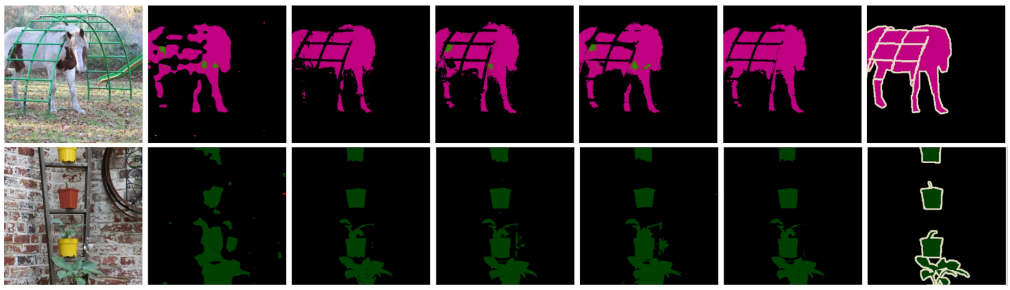
Similar to (Chambon et al., 2025), the VFM receives the normalized 448×448 image as input, and the extracted feature is upsampled to the original image resolution, corresponding to ×14 (Oquab et al., 2023) or ×16 (Siméoni et al., 2025; Tschannen et al., 2025) spatial scaling. Then, a linear classifier with a 1×1 convolution layer is trained on top of these upsampled features to predict per-pixel semantic labels or depth values. For VFM-specific methods (Couairon et al., 2025; Huang et al., 2025), we use their official implementations on each corresponding VFM as in (Chambon et al., 2025) and then perform linear probing. More details are included in **appendix**.

4.2 State-of-the-art Comparison

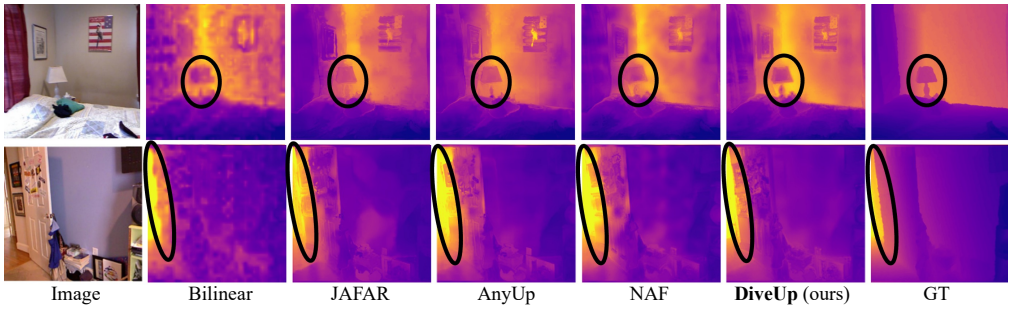
Comparison Across VFM-Agnostic Upsamplers. Tab. 1 compares DiveUp against VFM-agnostic upsamplers (Wimmer et al., 2026; Chambon et al., 2025) on Pascal VOC (Everingham et al., 2010) and NYUv2 depth (Silberman et al., 2012), reported individually for four structurally distinct ones held out of the training mixture and averaged over $m = 14$ backbones. DiveUp leads every column, by $+1.70$ mIoU and $+0.71$ δ_1 on the m-VFM average over the strongest baseline. The margin widens as the target moves away from the self-supervised ViTs these upsamplers are usually trained on: $+4.18$ mIoU on ConvNeXt-S against $+0.77$ on DINOv2-S. VFM-specific upsamplers are excluded here because an m-VFM average would require one training run per backbone; we compare against them in Tab. 2.

Comparison Across Tasks and Datasets. Tab. 2 shows the comparison on COCO-Stuff (Lin et al., 2014) and ADE20K (Zhou et al., 2017) and adds VFM-specific upsamplers, on DINOv2-S for comparability with prior work and on SigLIP-B, whose patch features carry far stronger spatial artifacts. Among VFM-agnostic methods, DiveUp is best in all eight cells. Against the VFM-specific ones the picture splits by backbone: on DINOv2-S, UPLiFT (Walmer et al., 2026) edges us on Pascal VOC (85.15 against 84.78) and ADE20K (43.02 against 42.71) while we lead COCO-Stuff and depth; on SigLIP-B we lead three of four (81.19, 60.99, 42.64), and only depth goes to UPLiFT. This is expected: a per-backbone upsampler can be fitted tightly to a clean feature space, and its advantage disappears once the target is noisy, at the cost of a separate training run for every backbone.

Qualitative Comparison of Upsampled Features. Fig. 3 visualizes the upsampled features of SigLIP-B using their first three PCA components, with a shared projection basis across all methods to ensure color comparability. We specifically select SigLIP-B for this visualization because language-image models notoriously exhibit severe spatial noise and high-norm patch artifacts compared to self-supervised ViTs. As observed in the raw low-resolution column, uniform regions (*e.g.*, the blue sky, water, and road surface) are dominated by heavy grid-like noise. Bilinear interpolation simply blurs this coarse patch grid. Strikingly, both NAF and AnyUp propagate this inherent source noise into uninformative background regions and severely fragment object contours, such as the aircraft's wings and the motorcycle's outline. In contrast, DiveUp suppresses background artifacts, rendering coherent flat regions. We attribute this success directly to our decoupled framework: the diverse training mixture prevents the upsampler from overfitting to the noisy spatial statistics of any single VFM family, while the teacher-guided COM field explicitly enforces boundary-aware spatial alignment. This qualitative structural superiority is supported quantitatively: DiveUp yields a $+3.44$ mIoU gain over baseline NAF on this backbone, and removing the COM term from DiveUp costs 1.80 points of the boundary-sensitive DAVIS $\mathcal{F}$ measure (App. S4.3).



(a) Results on Semantic Segmentation



(b) Results on Depth Estimation

Figure 4: Comparison of prediction results on semantic segmentation (a) in Pascal VOC and depth estimation (b) for different methods in NYUv2.

Qualitative Comparison on Prediction Results. To further demonstrate different methods, we show prediction results of upsamplers on semantic segmentation and depth estimation in Fig. 4. From Fig. 4, we can observe that, our DiveUp is able to produce better prediction results than other existing upsamplers, validating the importance and efficacy of learning from diverse VFMs.

4.3 Ablation Study

Efficacy of Geometric Guidance Across Architectures. Both multi-VFM mixed training and the teacher-guided COM field effectively mitigate the overfitting issues inherent in single-model self-reconstruction. To strictly isolate the contribution of the geometric guidance, Tab. 3 evaluates the COM loss independently by training the upsampler on a single backbone (App. S4.2 ablates the two jointly; their effects are not additive). Tab. 3 isolates what the loss contributes. Each block trains one upsampler on one backbone and applies it unchanged to probing backbones. Every gain is positive but varies by an order of magnitude: largest on CNNs farthest from the teacher's geometry (+12.77 ConvNeXt-S, +12.72 ResNet-50 when trained on SigLIP-B), and smallest on the better-localized CLIP-B (+1.47). Training on Franca (a ViT closer to the teacher) halves these gains (+4.25, +4.19, +1.26) without altering their rank. The consistent ranking suggests that backbones with a larger geometric mismatch from the teacher tend to benefit more, while the training backbone mainly affects the size of the gain.

Table 3: COM loss ablation.

Target VFM	w/o COM	w/ COM	Δ
Trained on SigLIP-B			
ConvNeXt-S	53.71	66.48	+12.77
ResNet-50	50.19	62.91	+12.72
Swin-S	58.94	67.78	+8.84
SigLIP-B	75.48	80.26	+4.78
MaskCLIP	71.81	76.45	+4.64
CLIP-B	75.34	76.81	+1.47
Trained on Franca			
ResNet-50	51.49	55.74	+4.25
ConvNeXt-S	55.52	59.71	+4.19
Swin-S	61.47	65.18	+3.71
SigLIP-B	77.97	79.99	+2.02
MaskCLIP	73.49	75.34	+1.85
CLIP-B	76.88	78.14	+1.26

Impact of Training Mixture Diversity. We measure how varied a mixture is by its CKA span (Kornblith et al., 2019), the spread of pairwise representational similarity among its backbones, and read generalization off the average over seven held-out VFMs. Fig. 5(a) reports both for three mix-

	Span ↑	Pascal VOC (mIoU ↑)		
		DINOv3-L†	ConvNeXt-S	**Holdout**
A	0.122	**88.60**	59.98	73.27
B	0.373	88.42	61.55	73.89
C	**0.591**	88.15	**64.04**	**74.98**

(a) Holdout: mean over 7 VFMs excluded from all training mixtures.

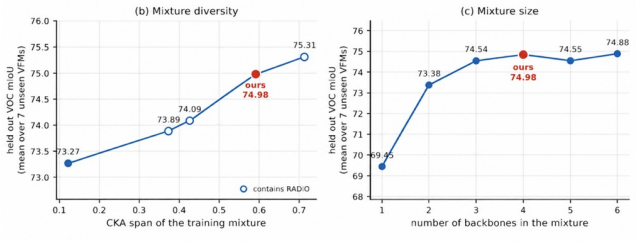


Figure 5: Training mixture ablation. Mixtures of increasing CKA (Kornblith et al., 2019) span. **A**: DINOv2-S, DINOv2-B^{reg}, DINOv2-S^{reg}, DINOv3-B; **B**: DINOv3-B, SigLIP, DINOv2-S, RADIOv2.5-B; **C**: DINOv3-B, SigLIP, Franca, PE-Core-S. *(b)* holdout accuracy against span; open markers contain RADIOv2.5-B. *(c)* holdout against mixture size.

tures: **A** draws all four checkpoints from the DINO family (span 0.122), **B** swaps two of them for SigLIP and RADIOv2.5-B (0.373), and our **C** instead takes SigLIP, Franca and PE-Core-S alongside DINOv3-B, covering four pretraining paradigms (0.591). Panel (b) plots the two against each other: they rise together, so span predicts transfer. Mixture A overfits to its own family, peaking on the in-distribution DINOv3-L (88.60 mIoU) but reaching only 59.98 on the unseen ConvNeXt-S. Mixture C concedes 0.45 mIoU on DINOv3-L and gains $+4.06$ on ConvNeXt-S, lifting the held-out average from 73.27 to 74.98. This indicates that diversity slightly reduces accuracy on the family already represented while improving transfer to architectures never seen. Panel (c) varies mixture size: the held-out average rises steeply to three backbones and is flat thereafter (74.54, 74.98, 74.55, 74.88 at three to six), showing no monotonic gain from simply adding backbones. The highest-span mixture we can build scores 75.31, but it contains RADIOv2.5-B, itself an agglomerative distillation of DINO, CLIP and SAM; training on it would weaken the "unseen architecture" claim, so we keep it out.

Due to space limitations, more extensive visualizations, ablation studies, and detailed experimental analyses are provided in the **appendix**.

5 Conclusion and Limitation

In this paper, we introduced DiveUp, a unified, encoder-agnostic framework for feature upsampling. DiveUp jointly trains a single shared upsampler on a diverse mixture of VFMs, reducing dependence on any individual source model and improving generalization to unseen architectures. Our analysis further shows that training-mixture diversity, quantified by CKA span, is an important factor for robust transfer. To provide additional spatial supervision, we introduced a local COM field as a model-agnostic relational representation that transfers geometry from a spatially aligned teacher without directly aligning heterogeneous feature spaces, with the multi-VFM reconstruction objective constraining semantic distortion. Extensive experiments demonstrate that DiveUp achieves state-of-the-art performance among VFM-agnostic upsamplers across diverse dense prediction tasks, with strong zero-shot generalization and no per-model retraining.

While DiveUp demonstrates robust zero-shot generalization across diverse architectures, one limitation lies in its dependence on the representational coverage of the training VFM mixture. Although a diverse set of training VFMs enables strong transfer across many unseen backbones, highly specialized or substantially out-of-distribution models may still be challenging when their feature representations differ markedly from those observed during training. A second limitation is that training requires access to a spatially well-localized teacher. Although the teacher is removed at inference and our ablations show robustness to teacher choice, the quality of the geometric supervision may depend on the availability of a suitable teacher for the target domain. This suggests that the generalization ability of the shared upsampler is partly determined by the breadth of the source representation space. Expanding the training mixture to include broader architectures, modalities, and domain-specific VFMs is therefore a natural direction for future work.

AI USE STATEMENT

Generative AI tools were used solely for brainstorming and language polishing. They were not used to generate data or experimental results, formulate mathematical claims, design experiments, implement methods, analyze results, or interpret findings. All AI-assisted text was carefully reviewed, revised, and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This paper presents DiveUp, a zero-shot feature upsampling framework designed to extract dense spatial representations from vision foundation models (VFMs). While our method itself is an architectural tool and does not explicitly collect new datasets or train foundation models from scratch, we recognize that the pre-trained VFMs (*e.g.*, CLIP, DINO) we utilize may inherit biases from their large-scale training data. Our upsampler relies on the native semantics of these models and therefore may inadvertently propagate these underlying biases into high-resolution downstream tasks. We encourage future practitioners to carefully audit the base VFMs for fairness and safety before deploying them in sensitive real-world applications.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our proposed DiveUp framework, we detail the experimental setup, training protocols, and hyperparameter configurations in Sec. 4.1 and Appendix. The precise architectural details of our default upsampler and the projection heads (including the semantic bottleneck) are provided in Appendix. Furthermore, we evaluate our method on standard, publicly available benchmarks. We will release the complete source code, pre-trained model weights, and evaluation scripts upon publication to facilitate future research.

REFERENCES

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

Loick Chambon, Paul Couairon, Eloi Zablocki, Alexandre Boulch, Nicolas Thome, and Matthieu Cord. Naf: Zero-shot feature upsampling via neighborhood attention filtering. *arXiv*, 2025.

Paul Couairon, Loick Chambon, Louis Serrano, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Jafar: Jack up any feature at any resolution, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Bing Fan, Yunhe Feng, Yapeng Tian, James Chenhao Liang, Yuewei Lin, Yan Huang, and Heng Fan. Prvql: Progressive knowledge-guided refinement for robust egocentric visual query localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5156–5165, 2025.

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

Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *ICLR*, 2024.

Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2013. doi: 10.1109/TPAMI.2012.213.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Haiwen Huang, Songyou Peng, Dan Zhang, and Andreas Geiger. Renovating names in open-vocabulary segmentation benchmarks. *Advances in Neural Information Processing Systems*, 37: 88410–88442, 2024.

Haiwen Huang, Anpei Chen, Volodymyr Havrylov, Andreas Geiger, and Dan Zhang. Loftup: Learning a coordinate-based feature upsampler for vision foundation models, 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics (ToG)*, 26(3):96–es, 2007.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

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

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.

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

Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *arXiv*, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Michael Ranzinger, Greg Heinrich, Pavlo Molchanov, Jan Kautz, Bryan Catanzaro, and Andrew Tao. Featsharp: Your vision model features, sharper. *arXiv*, 2025.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12490–12500, 2024.

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*, 2024.

Minseok Seo, Mark Hamilton, and Changick Kim. Upsample anything: A simple and hard to beat baseline for feature upsampling. *arXiv*, 2025.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

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*, 2025.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, R. Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim M. Alabdulmohsin, Lucas Beyer, and Xiao-Qi Zhai. Paligemma 2: A family of versatile vlms for transfer. *ArXiv*, abs/2412.03555, 2024.

Saksham Suri, Matthew Walmer, Kamal Gupta, and Abhinav Shrivastava. Lift: A surprisingly simple lightweight feature transform for dense vit descriptors. In *European Conference on Computer Vision*, pp. 110–128. Springer, 2024.

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*, 2025.

Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision*, pp. 367–385. Springer, 2024.

Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Spyros Gidaris, Elias Ramzi, Andrei Bursuc, and Yuki M. Asano. Franca: Nested matryoshka clustering for scalable visual representation learning. *arXiv preprint arXiv:2507.14137*, 2025.

Matthew Walmer, Saksham Suri, Anirud Aggarwal, and Abhinav Shrivastava. Uplift: Efficient pixel-dense feature upsampling with local attenders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41288–41298, 2026.

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv*, 2023.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. Anyup: Universal feature upsampling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *European Conference on Computer Vision*, pp. 453–469. Springer, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022.

APPENDIX

S1 PROPERTIES OF THE COM FIELD

Theoretical Properties of the COM Field. To understand what the COM field captures, consider an idealized image with a piecewise constant projected feature space: $z(q) = z_a$ on region A and $z(q) = z_b$ on region B (where $z_a \neq z_b$). Let $\mathcal{N}(p)$ be a symmetric $(2r{+}1) \times (2r{+}1)$ window centered at p. The field $\boldsymbol{\mu}(p)$ then exhibits two geometric properties.

Property 1 (Interior Vanishing). *If $\mathcal{N}(p) \subseteq A$, then $\boldsymbol{\mu}(p) \equiv \mathbf{0}$, for any $\tau > 0$ and under no condition on z_a.*
Proof sketch: Since $z(q) = z_a$ for all $q \in \mathcal{N}(p)$, the logits $\langle z_a, z_a \rangle / \tau$ are identical, yielding a uniform attention distribution $S(p, q) = 1/|\mathcal{N}(p)|$. By the spatial symmetry of the window, the unweighted sum of offsets cancels to zero.

Property 2 (Same-region centroid bias). *If $\mathcal{N}(p)$ intersects both A and B, and $\langle z_a, z_a \rangle > \langle z_a, z_b \rangle$, then*

$$\boldsymbol{\mu}(p) \;=\; \frac{w_A - w_B}{Z} \sum_{q \in A \cap \mathcal{N}(p)} (q - p), \tag{S1}$$

where $w_A = \exp(\langle z_a, z_a \rangle / \tau)$ and $w_B = \exp(\langle z_a, z_b \rangle / \tau)$ are unnormalized affinity weights, $w_A > w_B > 0$, and $Z = |A \cap \mathcal{N}(p)| w_A + |B \cap \mathcal{N}(p)| w_B$. Thus, $\boldsymbol{\mu}(p)$ is a positive multiple of the first spatial moment of the same-region support and points from p toward the centroid of $A \cap \mathcal{N}(p)$.
Proof. The inner-product assumption makes the unnormalized affinity weights piecewise constant, with weight w_A on $A \cap \mathcal{N}(p)$ and w_B on $B \cap \mathcal{N}(p)$. Window symmetry gives $\sum_{q \in \mathcal{N}(p)} (q - p) = \mathbf{0}$, hence $\sum_{B \cap \mathcal{N}(p)} (q - p) = -\sum_{A \cap \mathcal{N}(p)} (q - p)$. Substituting into Eq. (3) yields Eq. (S1), with coefficient $w_A - w_B > 0$. □

Corollary (Inward boundary bias). *If the boundary is locally straight inside the window, so that $A \cap \mathcal{N}(p)$ is a half-window, then the centroid of $A \cap \mathcal{N}(p)$ lies on the inward normal of A at p, and $\boldsymbol{\mu}(p)$ points strictly along that normal.*

We state the corollary separately because the direction is only normal to the boundary under that assumption. For a curved or irregular boundary, Eq. (S1) still holds exactly: the field points toward the centroid of the visible same-region support, although this direction need not coincide with the local boundary normal. This is the property the loss actually uses: it asks the student to reproduce a same-region centroid bias that vanishes in interiors, not one with a particular angle.

The Geometric Necessity of Feature Normalization. Property 1 guarantees that the COM field remains inert within flat regions, preventing structural hallucinations where the teacher provides no boundaries. Property 2 delivers the crucial localization signal, yet it fundamentally relies on the inner product condition $\langle z_a, z_a \rangle > \langle z_a, z_b \rangle$. Raw features frequently violate this: a neighboring region with an anomalously large magnitude can trivially invert the displacement vector. This failure mode mathematically explains the severe boundary artifacts caused by high-norm patches in vision-language models like SigLIP. Strict ℓ_2-normalization would make the condition automatic, since then $\langle z_a, z_a \rangle = 1 > \langle z_a, z_b \rangle$. We do not impose it. The GroupNorm inside $\phi(\cdot)$ standardizes each position across channels, which reduces the spread of feature magnitudes but does not bound the inner product, so it mitigates the failure mode rather than excluding it. We adopt it on empirical grounds: it is what the released configuration uses, and removing the bottleneck entirely costs 0.74 mIoU on held-out backbones (Tab. S3).

S2 EXTENDED QUANTITATIVE EVALUATIONS

S2.1 ZERO-SHOT GENERALIZATION ON NOISY BACKBONES

The main paper evaluates cross-VFM generalization on standard vision foundation models. Here we test the same upsampler, without retraining, against feature spaces that are substantially noisier or structured differently: multimodal features from PaliGemma2 (Steiner et al., 2024), text-aligned

Table S1: Zero-shot generalization across diverse and noisy VFMs, on PASCAL VOC (mIoU %), with no per-model retraining. None of these backbones is in the training mixture. Best and second best are **bold** and underlined.

Target VFM	Semantic Segmentation (mIoU %)				Absolute Gains	
	NAF	AnyUp-S	AnyUp-M	DiveUp (Ours)	vs. NAF	vs. AnyUp-M
ResNet-50 (He et al., 2016)	51.42	54.60	56.82	**59.79**	+8.37	+2.97
DFN-CLIP-B (Fang et al., 2023)	77.66	77.62	78.00	**79.11**	+1.45	+1.11
MaskCLIP-B (Zhou et al., 2022)	73.32	73.75	75.04	**76.54**	+3.22	+1.50
PaliGemma2 (Steiner et al., 2024)	79.58	79.64	80.24	**82.04**	+2.46	+1.80
ClearCLIP-B (Lan et al., 2024)	73.79	75.37	75.95	**76.73**	+2.94	+0.78

Table S2: **Plug-and-play across three upsampler designs**, PASCAL VOC (mIoU %). Every column is trained from scratch on the backbone it is evaluated on, with and without our COM-field term; no upsampler is transferred between backbones.

Target VFM	DiveUp (Ours)		AnyUp (Wimmer et al., 2026)		JAFAR (Couairon et al., 2025)		Δ		
	base	w/ COM	base	w/ COM	base	w/ COM	Ours	AnyUp	JAFAR
SigLIP-B	77.69	79.97	79.51	80.75	79.57	**81.32**	+2.28	+1.24	+1.75
CLIP-B	77.00	78.00	77.41	**78.07**	77.22	77.97	+1.00	+0.66	+0.75
MaskCLIP-B	73.99	74.93	73.98	74.63	74.94	**75.35**	+0.94	+0.65	+0.41
ConvNeXt-S	63.37	65.03	57.04	59.87	58.10	**65.68**	+1.66	+2.83	+7.58
ResNet-50	59.31	**60.35**	52.35	55.02	50.82	56.11	+1.04	+2.67	+5.29
Average	70.27	**71.66**	68.06	69.67	68.13	71.29	+1.38	+1.61	+3.16

spaces such as DFN-CLIP (Fang et al., 2023), and the CLIP variants MaskCLIP (Zhou et al., 2022) and ClearCLIP (Lan et al., 2024). None of these backbones appears in our training mixture.

Tab. S1 reports the result. Methods trained to reconstruct a single backbone (*e.g.*, NAF (Chambon et al., 2025), AnyUp-S (Wimmer et al., 2026)) lose ground as the target distribution moves away from what they were fit to, while DiveUp remains ahead on every backbone evaluated. The margin tracks how misaligned the target is: it is largest on ResNet-50, whose convolutional features are the furthest from any transformer in the mixture, and smallest on the CLIP variants, which the mixture partially covers through SigLIP. This is the behaviour the geometric objective is designed for, since the COM field supplies structure a backbone lacks rather than imposing a uniform prior.

S2.2 Plug-and-Play Robustness with Noisy VFMs

The main paper shows that our training objective can be attached to existing upsampler architectures. Here we check that this still holds when the target features are noisy, across three designs: our own upsampler, AnyUp (Wimmer et al., 2026) and JAFAR (Couairon et al., 2025). Each is trained from scratch on each target backbone separately, with its own objective otherwise unchanged, so nothing is transferred between backbones and the only difference between the two arms is the COM-field term.

Tab. S2 shows a gain on every backbone in all three designs, averaging $+1.38$, $+1.61$ and $+3.16$ mIoU. The gain is largest where the backbone is furthest from the teacher's geometry, the same ordering the main paper reports: on the two convolutional targets JAFAR gains $+7.58$ and $+5.29$, against $+0.41$ on MaskCLIP-B, which is already well localised. That ordering also explains why the three designs differ in how much they gain rather than in whether they gain: JAFAR starts lowest on the convolutional backbones (50.82 on ResNet-50) and so has the most to recover, while our upsampler starts highest (59.31) and gains least there. Because only the objective changes, the improvement is attributable to the geometric supervision rather than to capacity.

S3 Additional Ablation Studies

Two components of the COM field have free parameters: the width d of the projection the field is computed on, and the size $w \times w$ of the local window over which the self-affinity and the center of mass are taken. Tab. S3 ablates both; we then vary the teacher the field is read from.

Table S3: **COM field components. Left:** projection width d, scored as the mean over three held-out backbones (ResNet-50, ConvNeXt-S, Swin-S). **Right:** local window size, mean over 2-5 training seeds with one standard deviation.

Projection width d	Held-out mIoU (%)
4	**62.48**
16	62.11
64	62.47
256 (default)	62.40
1024 (no bottleneck)	61.66

Window size	SigLIP-B (mIoU)	DINOv2-S (mIoU)
3×3	78.96	84.30
5×5	**79.62**	84.43
7×7 (default)	79.61	**84.48**
9×9	79.42	84.42
11×11	79.28	84.37

Table S4: **Guidance teacher ablation**, PASCAL VOC (mIoU %). Both runs are identical to DiveUp except for the teacher the COM field is read from, and use the same probing protocol.

Probe backbone	DINOv3-L (Ours)	C-RADIO-v3-L (Ranzinger et al., 2024)	Δ
DINOv3-B	87.10	**87.15**	-0.05
SigLIP-B	**81.19**	80.79	$+0.40$
ConvNeXt-S	**63.55**	63.22	$+0.33$
ResNet-50	**59.79**	59.47	$+0.32$
Average	**72.91**	72.66	$+0.25$

Projection width. The COM field is computed on a low-dimensional projection of the teacher features rather than directly in the original feature space. Its performance is largely insensitive to the exact projection width: all tested bottlenecks from $d = 4$ to $d = 256$ fall within a 0.37 mIoU range on held-out backbones. More important than the precise width is the presence of a bottleneck itself. Computing the field at the teacher's full feature dimension reduces mIoU by 0.74, suggesting that high-dimensional teacher features introduce additional semantic interference into the geometric guidance. We therefore use $d = 256$ in all experiments.

Local window size. Performance follows an inverted U in w on both backbones. A 3×3 window is too small to establish a reliable affinity, so the centre-of-mass estimate is noisy; an 11×11 window reaches across object boundaries and averages in pixels from the far side of them, which blurs exactly the structure the field is supposed to sharpen. The peak sits between 5 and 7: on SigLIP-B the two are tied (79.62 against 79.61), and DINOv2-S separates them in favour of 7×7. We use $w = 7$. Entries are means over 2–5 training seeds; the seed spread is 0.00–0.06, roughly an order of magnitude below the differences between window sizes.

Choice of guidance teacher. The COM field is read off a single teacher, and the choice of that teacher is the one design decision the field inherits from outside our method. Tab. S4 swaps it. Replacing DINOv3-L with C-RADIO-v3-L, an agglomerative model distilled from DINO, CLIP and SAM, costs 0.25 mIoU on average, and the difference is not spread evenly: the two teachers are indistinguishable on DINOv3-B (-0.05), while the whole gap sits on the backbones furthest from the teacher's geometry (SigLIP-B $+0.40$, ConvNeXt-S $+0.33$, ResNet-50 $+0.32$ in favour of DINOv3-L). These margins are small enough that we read them as a robustness result rather than a ranking: the method needs a well-localised teacher, not this particular one.

S4 Deep Dive into Geometric Representations

S4.1 Visualizing the Entropy Gate and the COM Field

To examine why the gain concentrates on structurally noisy backbones, we visualise the two intermediate quantities the geometric objective is built from: the local entropy map and the COM field. Both are computed on a low-dimensional projection of the features over a 7×7 window. Within that window the feature similarities are softmax-normalised into a distribution; its entropy H measures how sharply the window commits to a direction, and the first moment of the same distribution gives the local centre-of-mass displacement, the COM field.

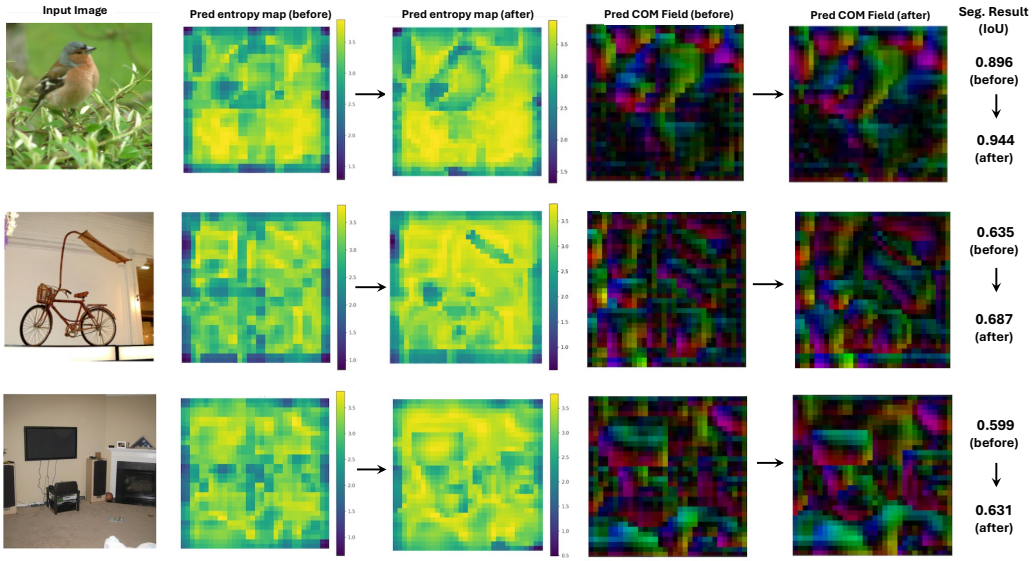


Figure S1: Entropy maps and COM fields, before and after training with the COM objective. We visualize the intermediate local entropy maps and COM fields generated by a noisy backbone (ClearCLIP-B), compared with those generated after integrating DiveUp's COM Field guidance. The guidance suppresses background interference and spatial uncertainty, achieving more localized geometric structures. This improved location alignment corresponds to the quality of the final segmentation results (see IoU gains).

Fig. S1 compares these quantities before and after training with the COM objective.

Before (baseline). The target VFM's native features are spatially ambiguous. Entropy stays high and undifferentiated across object interiors and background alike, so there is no clean separation to key on, and the corresponding COM field carries directional vectors that point inconsistently, including inside regions that contain no boundary.

After. Supervised by the COM field of a single spatially-aligned teacher, the student's own entropy map drops inside objects and stays high only near contours. The predicted entropy maps effectively suppress internal and background noise, producing clearer object contours. Similarly, the predicted COM fields (visualized via HSV maps) align more accurately with actual object boundaries and show fewer background artifacts.

These refined geometric representations correct the location misalignments inherent in the source VFM. This structural improvement contributes to the enhanced downstream segmentation performance, which is consistent with the absolute IoU gains observed in the comparisons.

S4.2 Ablating the Mixture and the COM Term Together

Sec. 4.3 isolates the COM term by training on a single backbone, which is the setting in which its contribution is largest and easiest to read. Here we vary both mechanisms at once.

Table S5: **The two mechanisms are not additive.** Held-out average over the four backbones the two settings share (MaskCLIP, ResNet-50, Swin-S, ConvNeXt-S). On full-image mIoU either mechanism alone recovers most of what both recover together; inside a ±2px boundary band they combine.

Training mixture	**Full mIoU**		**Boundary band (±2px)**	
	w/o COM	w/ COM	w/o COM	w/ COM
Single backbone (SigLIP-B)	58.66	**68.41**	40.97	52.89
Four-VFM mixture *(ours)*	66.83	66.94	52.83	**53.98**

Read on full-image mIoU the two mechanisms substitute for each other: the mixture alone recovers $+8.17$ over the unguided single-backbone model, the COM term alone $+9.74$, and both together $+8.27$ – no more than either (all differences from unrounded values). This is what we should expect if they correct the same defect, and it is why the main paper ablates them separately rather than reporting a decomposition that does not decompose. The picture changes inside a $\pm$2px boundary band, which covers 3.8% of pixels against 90.2% for object interiors: there the two combine, $+13.01$ against $+11.92$ and $+11.86$ for each alone.

The row that looks strongest on full-image accuracy is also the one we do not ship. The single-backbone model with COM reaches 68.41, but it buys that by contracting its feature space, and on DAVIS-2017 propagation the ordering reverses on every backbone we evaluated: $+4.90$ $\mathcal{J}\&\mathcal{F}$ on ConvNeXt-S, $+5.20$ on ResNet-50, $+5.20$ on CLIP-B and $+6.80$ on DINOv3-B in favour of the mixture. App. S4.4 measures that contraction and shows which intervention removes it.

Tab. S6 reports the COM term on its own in the configuration we deploy, over all 14 backbones. Its effect on full-image mIoU is $+0.20$ and positive on 13 of 14; scored inside the boundary band the same comparison gives $+1.25$ and is positive on all 14, with the gain largest on the agglomerative and language–image models ($+1.49$, $+1.26$) and smallest on the DINO family ($+1.19$), whose features are already well localised. The convolutional backbones separate the two readings most sharply: $+0.02$ on full mIoU against $+1.13$ at the boundary.

Table S6: **COM term in the deployed configuration.** The four-VFM mixture trained with and without the COM field, probed on all 14 backbones and grouped by architecture family.

Evaluation backbones	**Full mIoU**			**Boundary band ($\pm$2px)**		
	w/o COM	w/ COM	Δ	w/o COM	w/ COM	Δ
DINO family *(4)*	85.89	86.07	+0.19	70.77	71.96	+1.19
Language–image *(5)*	78.04	78.31	+0.27	62.91	64.17	+1.26
Agglomerative *(2)*	87.57	87.88	+0.30	72.06	73.55	+1.49
CNN / hierarchical *(3)*	63.71	63.73	+0.02	49.96	51.09	+1.13
All 14	78.57	**78.77**	+0.20	63.69	**64.93**	+1.25

S4.3 Boundary Quality on Video Propagation

Sec. 4 reports a boundary measure on DAVIS-2017. We use the standard semi-supervised protocol: the first-frame mask is propagated through the clip by nearest-neighbour matching in upsampled feature space, and the prediction is scored with the benchmark's two measures – $\mathcal{J}$, the region IoU, and $\mathcal{F}$, the F-measure between the predicted and ground-truth object boundaries after a morphological dilation. $\mathcal{F}$ is the boundary-sensitive half of the pair, which is why we report it here. Tab. S7 ablates the COM term in the deployed mixture under this protocol.

Table S7: **DAVIS-2017 semi-supervised propagation**, the four-VFM mixture with and without the COM term. $\mathcal{F}$ is the boundary measure. Both arms use the same propagation protocol and differ only in the training objective.

Target VFM	$\mathcal{J}\&\mathcal{F}$			$\mathcal{F}$ **(boundary)**		
	w/o COM	w/ COM	Δ	w/o COM	w/ COM	Δ
DINOv2-S	69.00	70.00	+1.00	71.30	72.50	+1.20
DINOv3-B	68.50	68.60	+0.10	70.90	71.10	+0.20
SigLIP-B	62.70	64.10	+1.40	65.20	67.00	+1.80
PE-Core-S	51.60	51.30	-0.30	53.40	53.30	-0.10
CLIP-B	59.70	61.50	+1.80	62.10	64.00	+1.90
ResNet-50	39.40	41.30	+1.90	37.30	39.30	+2.00
ConvNeXt-S	38.10	39.60	+1.50	37.20	38.70	+1.50
Average	55.57	**56.63**	+1.06	56.77	**57.99**	+1.21

S4.4 Why the Mixture Is Needed for the COM Term to Be Safe

The mixture is introduced in Sec. 3 as a way to generalize across backbones. It has a second role, which a video-propagation evaluation makes visible: it is also what keeps the COM term from degrading the feature space it is applied to.

The phenomenon. On DAVIS-2017 semi-supervised propagation, adding the COM term to an upsampler trained on a *single* backbone lowers $\mathcal{J}$ on all six backbones we tried, while adding it to the four-VFM mixture improves the average $\mathcal{J}$ and $\mathcal{F}$, with positive gains on six of seven evaluated backbones and only a small decrease on PE-Core-S (Tab. S7). The same loss and evaluation therefore behave very differently across the two training settings. Segmentation accuracy does not show this: the single-backbone setting gains $+5.11$ mIoU on VOC over the same seven held-out backbones.

The cause. Measuring foreground-query similarities on DAVIS sequences, the single-backbone COM run raises same-class and different-class cosine similarity *together*, so class separability falls by 22%; under the mixture it does not move (Tab. S8, left). Propagation is recursive, so a narrower margin occasionally sends a label to a semantically similar but spatially unrelated region, and the error is locked in for the rest of the clip. This is visible in the failure mode: missed detections are unchanged, and the false positives change character – without the term 97% of them lie within 8px of a true boundary, with it only half do.

Why the term should contract at all follows from Property 1. The ± 2px boundary band covers 3.8% of pixels and object interiors 90.2%, so 90% of the loss is computed in interiors, where the teacher's displacement is small but not zero. The cheapest way to reproduce a small displacement everywhere is to flatten the local softmax, which by Property 1 means pulling neighbouring features together – a contraction. Reconstructing a single target space does not penalise it, because the contracted student still fits that one space; sampling a different target VFM each step does, because no contracted solution fits four spaces at once.

A causal test. Gating the loss to the high-entropy band, which removes exactly the interior terms, recovers 87% of the lost separability and improves DAVIS on all three backbones tested, while the VOC gain is preserved rather than traded away ($+4.78 \rightarrow +5.60$ on SigLIP-B, $+1.47 \rightarrow +2.46$ on CLIP-B). One intervention moves the proposed cause and the observed effect together.

Two consequences for how the rest of this paper should be read. First, the large single-backbone numbers in Tab. 3 should not be extrapolated to the deployed model: part of that gain is accompanied by a contraction that the diverse mixture substantially mitigates. Second, the mixture is not only a generalization device; it also helps keep the geometric term from degrading semantic separability.

Table S8: **Feature-space contraction under the COM term. Left:** cosine similarity and class separability of foreground queries on DAVIS-2017. **Right:** the resulting change in propagation quality, averaged over the backbones evaluated in each setting.

Training setting	DAVIS foreground queries			Propagation	
	same-class cos	diff-class cos	**separability**	$\Delta\mathcal{J}$	$\Delta\mathcal{F}$
Single backbone, w/o COM	0.7931	0.4977	0.2954	–	–
Single backbone, w/ COM	0.9209	0.6915	0.2295 (-22%)	-2.28	$+0.43$
Single backbone, w/ COM + entropy gate	0.8763	0.5892	0.2871 (-3%)	$+1.68^{*}$	$+3.69^{*}$
Mixture, w/o COM	0.8588	0.5558	0.3030	–	–
Mixture, w/ COM *(ours)*	0.8509	0.5382	0.3127 ($+3\%$)	$+1.06$	$+1.21$

*ConvNeXt-S, the backbone where the gate was evaluated; the other two move from -4.73 to -3.75 (DINOv3-B) and from -2.01 to -0.21 (CLIP-B).

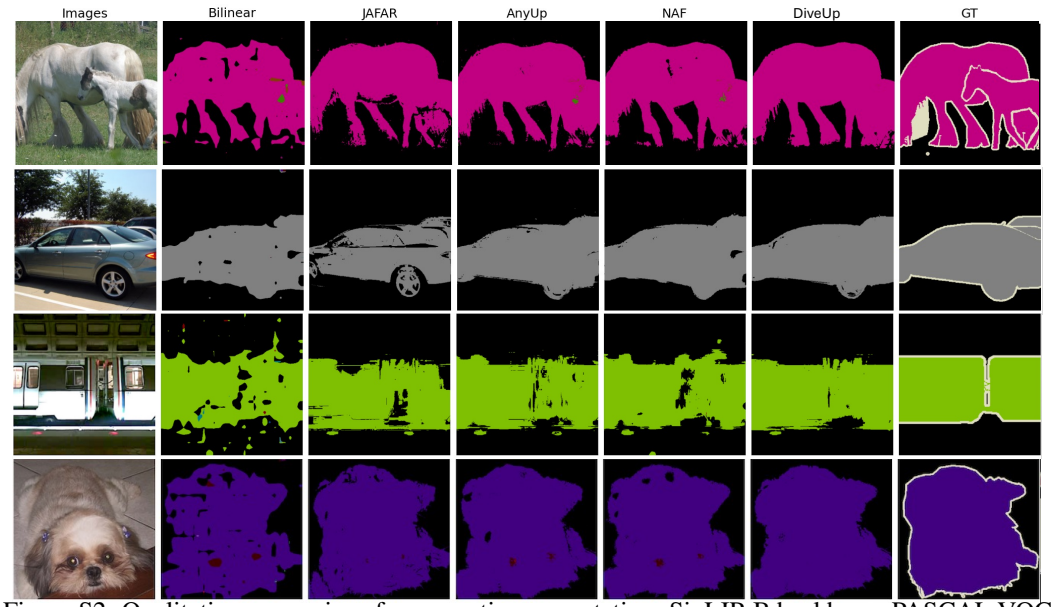


Figure S2: Qualitative comparison for semantic segmentation. SigLIP-B backbone, PASCAL VOC. Compared with both VFM-agnostic and VFM-specific upsamplers, DiveUp suppresses semantic bleeding and recovers sharper object boundaries.

S5 Comprehensive Qualitative Results

We provide qualitative comparisons against baseline methods on semantic segmentation and dense depth estimation to support the quantitative results.

S5.1 High-Resolution Semantic Segmentation

Fig. S2 visualises semantic segmentation produced from SigLIP-B features. As a vision-language model, SigLIP is spatially misaligned and carries high-norm artifacts, which makes it the setting where the geometric term matters most.

Baseline upsamplers (JAFAR, AnyUp, NAF) do not fully correct these artifacts: object boundaries stay blurred and semantics leak into the background. DiveUp supervises the same architecture with a teacher-guided COM field, which transfers the local geometry of a spatially-aligned teacher without directly aligning heterogeneous semantic feature spaces. This results in more precise object boundaries during probing and substantially fewer artifacts in object interiors.

S5.2 Dense Depth Estimation

Fig. S3 compares dense depth estimation from DINOv2-S features. DINOv2 already provides strong, well-localized patch representations, so this is the opposite regime from the previous figure: standard upsampling still introduces jagged or blocky edges at depth discontinuities, but the headroom for geometric correction is small. Aligning upsampling with the COM field preserves these discontinuities. Consistent with that reading, the quantitative gain from the COM term is largest on the structurally noisiest backbones and negligible on well-aligned ones; the loss supplies geometry a backbone lacks rather than imposing a uniform prior.

S6 Detailed Architecture and Evaluation Protocols

S6.1 Upsampler Architecture

Our upsampler $\mathcal{U}_\theta$ adopts the lightweight, VFM-agnostic design of NAF (Chambon et al., 2025): a cross-scale neighborhood attention that fuses high-resolution spatial structure from the image with

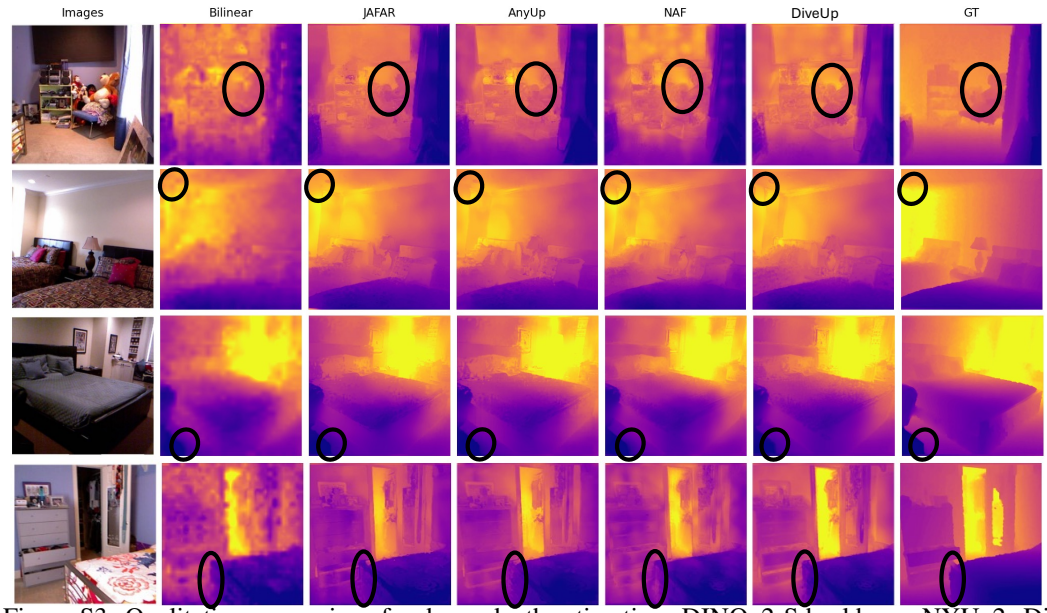


Figure S3: Qualitative comparison for dense depth estimation. DINOv2-S backbone, NYUv2. DiveUp preserves depth discontinuities and complex object boundaries.

low-resolution VFM semantics. We keep the architecture fixed throughout the paper so that every reported gain is attributable to the training mixture and the COM field rather than to capacity. Algorithm 1 gives the forward pass; the configuration is embedding width 256, 9×9 neighborhood attention, 4 attention heads, 4 RoPE heads, and 0.66M parameters in total.

Why the guidance stem is two parallel kernels. The image encoder is not a single convolutional stack. It runs a 1×1 branch and a 3×3 branch in parallel, each emitting $C/2$ channels, and concatenates them (Algorithm 1, line 2). The 1×1 branch responds to per-pixel color, which is what makes upsampled features look sharp; the 3×3 branch sees a local neighborhood and carries the context needed to decide whether a color edge is an object boundary or a texture. Relying on the 1×1 path alone produces visibly sharper maps that follow high-frequency RGB texture into regions where the semantics do not change, which is why sharper output is not automatically better for downstream probing.

Queries, keys, and values. Guidance is extracted once from the high-resolution image and reused for both queries and keys, so the attention pattern is a function of the image alone and never of the VFM. Queries are the RoPE-encoded guidance at the output grid; keys are the same tensor average-pooled to the VFM's grid, which keeps query and key in one coordinate frame. The values are the raw VFM features F_{lr}, passed through with *no* linear projection. This is the property that makes the upsampler VFM-agnostic at inference: the output lives in the target VFM's own feature space, so a model trained on one mixture can be applied to a backbone it has never seen without any per-backbone adaptation.

Resolution alignment. Training uses I_{hr} at 512×512 and its $2\times$ downsampled counterpart I_{lr}, i.e. a $\times 2$ ratio. Inference is not tied to that ratio. Queries are placed at whatever output grid is requested and RoPE encodes continuous coordinates, so the same weights extrapolate to the $\times 14$ and $\times 16$ factors that patch-based VFMs require, in a single forward pass and with no retraining or tiling. All results in this paper are produced this way.

Single-stage inference. NAF reaches its published numbers with an additional Stage-2 refinement pass. Every DiveUp number we report is Stage-1 only. Per-forward-pass cost (parameters, FLOPs, peak memory) is therefore comparable to NAF's Stage-1; where we claim efficiency, we mean it at the level of the deployment paradigm, in that N per-backbone upsamplers are replaced by one model that transfers zero-shot, not that a single forward pass is cheaper.

Algorithm 1 Architecture of the Upsampler $\mathcal{U}_\theta$

Input: Low-resolution VFM features $F_{lr} \in \mathbb{R}^{h\times w\times C}$, High-resolution image $I_{hr} \in \mathbb{R}^{H\times W\times 3}$, output grid (H_o, W_o), neighborhood size k=9

// 1. Spatial Guidance Extraction (image only; independent of the VFM)
$G = \text{Concat}\big(\text{Conv}_{1\times1}(I_{hr}),\ \text{Conv}_{3\times3}(I_{hr})\big)$ // $C/2$ channels each: detail ‖ context
$G = \text{AdaptiveAvgPool}(G,\ (H_o, W_o))$ // resample to the requested output grid
$G_{RoPE} = \text{Apply2DRoPE}(G)$ // continuous coordinates; any upsampling factor
// 2. Query, Key, and Value Formulation
$Q_{hr} = G_{RoPE}$ // high-resolution spatial queries
$K_{lr} = \text{AdaptiveAvgPool}(G_{RoPE},\ (h, w))$ // same frame, VFM grid
$V_{lr} = F_{lr}$ // raw VFM semantics, no projection
// 3. Cross-scale Neighborhood Attention
$\hat{F}_{hr} = \text{NeighborhoodAttention}(Q_{hr}, K_{lr}, V_{lr},\ k)$ // each query attends to a $k \times k$ local window
return $\hat{F}_{hr} \in \mathbb{R}^{H_o\times W_o\times C}$ // same channel space as F_{lr}

S6.2 The COM Projection Heads

The COM field is computed on a projection of the features rather than on the features themselves, and because the models in a mixture do not share a channel width, that projection cannot be a single layer. This subsection states exactly what is instantiated, since the student heads are the only learnable parameters outside the upsampler; the teacher head is frozen.

A head is a GroupNorm followed by a 1×1 convolution without bias, $\phi_C : \mathbb{R}^C \to \mathbb{R}^{d_\phi}$, with $d_\phi = 256$ throughout. The GroupNorm uses a single group, so it normalises over all channels at a position rather than over subsets of them. Student and teacher use *separate* heads: one ϕ^s_C per distinct student channel width present in the mixture, and one $\phi^t_{C'}$ per teacher. Our mixture contains DINOv3-B, SigLIP-B and Franca at C=768 and PE-Core-S at C=384, so training instantiates two student heads, shared by the backbones of equal width, and one teacher head for DINOv3-L. A backbone therefore does not own a private head; backbones of the same width share one. The alternative – a head per backbone – is implemented but disabled in the released configuration.

The student heads are trained jointly with the upsampler, while the teacher head is frozen. All projection heads are *discarded at inference*: they take part only in $\mathcal{L}_{\text{guide}}$, never in $\mathcal{U}_\theta$, so the deployed model has no per-VFM parameters and an unseen backbone needs none. The projection width is ablated in Tab. S3; what matters is that a bottleneck exists rather than its exact size.

Two symbols in Sec. 3.1 and this subsection must not be confused: C_M is the sampled model's native channel width, which the upsampler preserves, and d_ϕ is the width of the COM projection, which is 256 regardless of C_M.

S6.3 Definition of CKA Span

Sec. 4.3 orders training mixtures by their CKA span. The quantity is defined as follows. For two models M_i, M_j we extract one global feature vector per image – the `[CLS]` token where the backbone has one, and the mean over spatial tokens otherwise – on N=300 images sampled at regular intervals from the PASCAL VOC validation split at 448^2 resolution, giving matrices $X \in \mathbb{R}^{N\times D_i}$ and $Y \in \mathbb{R}^{N\times D_j}$. After centring each column, linear CKA (Kornblith et al., 2019) is

$$\text{CKA}(M_i, M_j) \;=\; \frac{\|X^\top Y\|_F^2}{\|X^\top X\|_F\ \|Y^\top Y\|_F}, \tag{S2}$$

which is defined for $D_i \neq D_j$, so backbones of different width need no projection. The span of a mixture is the range of its pairwise values,

$$\text{Span}(\mathcal{M}) \;=\; \max_{i\neq j} \text{CKA}(M_i, M_j) \;-\; \min_{i\neq j} \text{CKA}(M_i, M_j), \tag{S3}$$

over the $\binom{K}{2}$ pairs of a K-model mixture; for K=4 this is six values. We use the range rather than the mean because a mixture can have a low mean similarity while still containing no pair that is genuinely far apart, and it is the presence of distant pairs that the ablation is about. Span is a

Algorithm 2 Linear Probing for Semantic Segmentation

Input: High-resolution image $I \in \mathbb{R}^{H\times W\times 3}$, Ground Truth Mask $M_{gt} \in \mathbb{R}^{H\times W}$
Input: Frozen VFM Backbone $\mathcal{B}$, Frozen Upsampler $\mathcal{U}_\theta$
Input: Trainable 1×1 Conv Weight $W_{seg} \in \mathbb{R}^{N_{cls}\times C\times 1\times 1}$
1: // 1. Feature Extraction (No Gradients)
2: Extract low-res semantics: $F_{lr} = \mathcal{B}(I)$ // Frozen backbone forward
3: Predict high-res features: $\hat{F}_{hr} = \mathcal{U}_\theta(I, F_{lr})$ // Frozen upsampler forward
4: Detach from computation graph: $\hat{F}_{hr} = \text{StopGradient}(\hat{F}_{hr})$
5: // 2. Linear Classification
6: Predict pixel-wise logits: $S = W_{seg} * \hat{F}_{hr}$ // $S \in \mathbb{R}^{N_{cls}\times H\times W}$
7: Compute objective: $\mathcal{L}_{CE} = \text{CrossEntropy}(S, M_{gt})$
8: // 3. Optimization
9: **return** $\nabla_{W_{seg}}\mathcal{L}_{CE}$ // Update ONLY the linear classifier

Algorithm 3 Linear Probing for Dense Depth Estimation

Input: High-resolution image $I \in \mathbb{R}^{H\times W\times 3}$, Ground Truth Depth $D_{gt} \in \mathbb{R}^{H\times W}$
Input: Frozen VFM Backbone $\mathcal{B}$, Frozen Upsampler $\mathcal{U}_\theta$
Input: Trainable 1×1 Conv Weight $W_{dep} \in \mathbb{R}^{256\times 2C\times 1\times 1}$ // 256 depth bins
1: // 1. Feature Extraction (No Gradients)
2: Extract dense semantics: $F_{lr} = \mathcal{B}(I)$
3: Predict high-res features: $\hat{F}_{hr} = \mathcal{U}_\theta(I, F_{lr})$ // Frozen upsampler forward
4: Detach from computation graph: $\hat{F}_{hr} = \text{StopGradient}(\hat{F}_{hr})$
5: // 2. Bin-based Depth Classification
6: Duplicate features: $F_{in} = \text{Concat}(\hat{F}_{hr}, \hat{F}_{hr})$ // JAFAR places `[CLS]` here; ours has none
7: Predict bin logits: $S = W_{dep} * F_{in}$ // $S \in \mathbb{R}^{256\times H\times W}$
8: Normalize probabilities: $P = \frac{\text{ReLU}(S)+0.1}{\sum(\text{ReLU}(S)+0.1)}$ // Soft depth assignment
9: Expected depth: $D = \sum_{k=1}^{256} P_k \cdot \text{bin}_k$ // Linear combination of bins
10: // 3. Optimization
11: Compute objective: $\mathcal{L}_{depth} = \text{SigLoss}(D, D_{gt}) + 0.5\,\text{GradientLoss}(D, D_{gt})$
12: **return** $\nabla_{W_{dep}}\mathcal{L}_{depth}$ // Update ONLY the linear regressor

descriptive statistic of a mixture, not a quantity we optimise: the mixtures compared in Sec. 4.3 were chosen to vary architecture and pretraining paradigm, and their spans were measured afterwards.

S6.4 Linear Probing Evaluation Protocols

To evaluate the representation quality of the upsampled features $\hat{F}_{hr}$ we follow the linear probing protocol of prior work (Couairon et al., 2025; Chambon et al., 2025). Both the VFM backbone $\mathcal{B}$ and the trained upsampler $\mathcal{U}_\theta$ are frozen; the probing head is a single 1×1 convolution with no non-linearity and no multi-layer decoder, so the evaluation rests entirely on a pixel-wise linear objective.

Shared settings. Images are resized so that the short side is 448 and centre-cropped to 448×448; segmentation masks receive the same geometry with nearest-neighbour interpolation. Probes are optimised with AdamW at 5×10^{-4} under a cosine schedule, batch size 2, for 20 epochs on Pascal VOC and NYUv2 and 5 epochs on the larger ADE20K and COCO-Stuff. During probe training we apply a random horizontal flip with probability 0.5 to the image and its target jointly. For segmentation this matches the released protocol of (Chambon et al., 2025) parameter for parameter; we changed none of its training hyper-parameters.

The training pipelines for Semantic Segmentation and Dense Depth Estimation are detailed in Algorithm 2 and Algorithm 3 respectively.

Notes on the depth protocol. Three properties of Algorithm 3 need stating: First, the head is $2C$ wide. JAFAR fills the second half with the broadcast `[CLS]` token; our upsampler emits no `[CLS]`, so we duplicate the dense features instead (line 6). The two are the same function class,

$(W_1 + W_2)x$ versus Wx, but the duplicated form carries twice the parameters and therefore moves at roughly twice the effective learning rate. We verified that the head form is almost immaterial (0.07 δ_1 at 10 epochs) while the epoch budget is not ($+2.8$ from 10 to 20). Second, we could not reuse the reference implementation. The depth-probing script released with (Chambon et al., 2025) does not run: it imports a `DepthLoss` symbol absent from the accompanying `src/loss.py`, and the repository ships no NYUv2 dataset module. We therefore reimplemented the probe and trained every method in the column ourselves, baselines included, with the objective on line 12 throughout. Third, our protocol is not the Probe3D (El Banani et al., 2024) protocol that (Chambon et al., 2025) and AnyUp (Wimmer et al., 2026) report against. Probe3D trains the depth probe on the $\sim$ 30k-frame NYU-GeoNet annotations and evaluates on the 654 official test indices; we train and evaluate on the $1,449$ densely labelled frames of NYUv2. The training set is therefore roughly $40\times$ smaller, which accounts for the gap between our absolute δ_1 values and those in the literature. Within our table every method sees the identical protocol, so the ordering is meaningful even though the absolute level is not comparable across papers.