THE COUPLING FORM OF POSITIONAL ENCODING DETERMINES MULTI-TEACHER GENERALIZATION IN FEATURE UPSAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-attention feature upsamplers increase the spatial resolution of vision foundation model (VFM) features. Recent upsamplers such as NAF build queries and keys from two sources: image content encoded by a lightweight image encoder, and pixel positions supplied by a rotary position embedding (RoPE). Training such an upsampler with multiple VFMs as teachers is a natural way to improve its transfer to unseen VFMs, yet it has been reported to yield no clear gains, and the cause is not understood. We show that the cause lies in how position is coupled with content. RoPE rotates queries and keys, so position multiplies content. Different teachers require different shapes of the spatial kernel, but under RoPE only the shared image encoder can produce this difference, so the gradients of the teachers conflict within the first 100 training steps. This conflict is resolved not by optimization that modifies the gradients, but by replacing RoPE with a relative position bias. This bias adds position to the attention score as a separate term and learns the kernel shape instead of the image encoder. By building every scheme on NAF and swapping only the positional encoding, on Cityscapes semantic segmentation, adding a second teacher VFM causes a 0.83 mIoU drop for RoPE compared to its best single-teacher baseline, whereas the schemes that add position to the attention score gain up to 0.57 mIoU. An upsampler that learns from multiple VFMs should add position to the attention score instead of multiplying it with content.

1 INTRODUCTION

Vision foundation models (VFMs) such as DINOv3 (Siméoni et al., 2026) and CLIP (Radford et al., 2021) produce representations that serve a wide range of visual tasks. The patch embedding of the Vision Transformer (Dosovitskiy et al., 2021), however, leaves these representations at a coarse spatial resolution, which forms a bottleneck in dense prediction tasks such as segmentation and depth estimation. Feature upsamplers that increase the resolution through cross-attention have drawn interest as a remedy (AnyUp; Wimmer et al., 2026). Methods such as NAF (Chambon et al., 2026) and RaysUp (Ding et al., 2026) pass the input image through a lightweight convolutional encoder to produce queries and keys, and use the coarse VFM features as values. Such an upsampler is trained on image pairs of different resolutions, with a self-reconstruction loss that aligns the upsampled features of the low-resolution image with the features of the high-resolution one. Since queries and keys are derived from the image rather than the features, the upsampler is VFM-agnostic at inference and transfers to unseen VFMs without retraining.

NAF interprets this cross-attention as an extension of classical joint bilateral filtering. It treats the pattern of attention weights with which each output position gathers its neighbors as a kernel with two components: a local spatial kernel, given by RoPE (Su et al., 2024), and an adaptive content kernel, given by the similarity of image features. Without a positional encoding, the attention score depends only on that similarity, not on distance. The positional encoding is therefore the only element that gives the kernel its spatial shape.

Training typically uses a single VFM as the teacher, such as DINOv2 or DINOv3. However, an upsampler trained on a single VFM tends to overfit to the spatial biases and artifacts specific to that model (DiveUp; Liu & Fan, 2026). A natural way to prevent this is multi-teacher training on several VFMs simultaneously. Yet prior work reports that mixing teachers fails to outperform the best single teacher (Chambon et al., 2026, Tab. 10) , and DiveUp reports the same for the multi-VFM version

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.

of AnyUp (Liu & Fan, 2026, Tab. 3). The only explanation offered by DiveUp is that such mixing dilutes the capacity of the upsampler, leaving the underlying cause an open question.

This paper argues that in multi-teacher training with a shared reconstruction objective, the coupling form of positional information and content features determines success or failure. Earlier methods, NAF and RaysUp among them, have consistently adopted RoPE (Rotary Position Embedding). RoPE multiplies queries and keys by a coordinate-dependent rotation matrix, and AnyUp and RaysUp only change how the rotation angle is set or how the coordinates are taken. All of these methods multiply positional information into the content features (multiplicative coupling). We turn instead to additive coupling, which, as in T5 (Raffel et al., 2020) and Swin Transformer (Liu et al., 2021b), adds positional information to the computed attention score as an independent term. How the coupling form itself affects multi-teacher training, however, has been overlooked entirely (§2).

To test this claim, we build every scheme on the same NAF architecture, swap only the positional encoding (§3), and derive three predictions: under multiplicative coupling, (1) a second teacher brings a smaller gain than under additive coupling, (2) the teachers' gradients conflict inside the shared convolutional encoder in early training, and (3) gradient projection does not close this gap.

We evaluate downstream performance on seven VFMs, five of them unseen during training (§4). On Cityscapes semantic segmentation, adding a second teacher VFM leaves RoPE below its best single teacher, even for a variant with learnable frequencies, while the additive schemes gain up to 0.57 mIoU (Figure 1, §4.2). On DAVIS video object segmentation, the gap between the two coupling forms widens: both RoPE variants fall below the additive ones by an average of 3.5 $\mathcal{J}\&\mathcal{F}$ (§4.2).

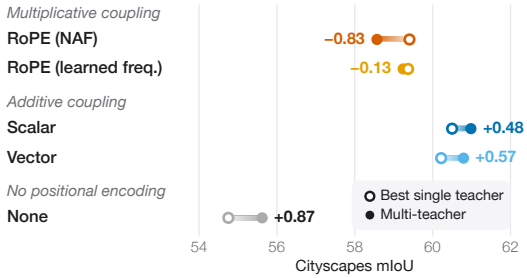


Figure 1: Cityscapes mIoU with the best single teacher and with both DINOv3 and SigLIP2 (mean over the seven evaluation VFMs). Numbers give the gain from the second teacher.

To explain the difference between the two coupling forms, we analyze the per-teacher gradients in early training (§4.3). Under RoPE, the teachers' demands for different spatial kernels flow into the convolutional encoder that produces queries and keys. The cosine similarity between the two teachers' gradients falls steadily over the first hundred steps and turns negative. Under the additive scheme, the per-teacher gradients keep the same direction. Position stays a separate term, and the convolutional encoder receives no conflicting signal. Late in training, the cosine is negative under both forms, because the two gradients sum to zero at a stationary point of the training loss; only the early conflict is specific to multiplicative coupling (§4.3).

If the gradient conflict under RoPE was merely an optimization problem of the kind known in multi-task learning, gradient projection (PCGrad; Yu et al., 2020) would resolve it (§2). Yet applying PCGrad to RoPE recovers only a fifth of the drop (§4.4). Multi-teacher training breaks down because multiplicative coupling lets gradient conflict develop, and additive coupling removes this failure.

In summary, this paper makes the following contributions:

- **Coupling form as the axis.** Controlled experiments show that negative transfer in multi-teacher training originates in the coupling form of positional information and content features.
- **The mechanism of gradient conflict.** Per-teacher gradients in early training locate the failure: multiplicative coupling puts the teachers in conflict inside the convolutional encoder, whereas additive coupling does not.
- **A design rule.** Replacing the positional encoding with a relative position bias, and adding nothing else, turns multi-teacher training into a gain.

2 Related Work

Feature upsampling. Learned feature upsamplers began with task-specific methods (Wang et al., 2019; Lu et al., 2022; Liu et al., 2023), moved to methods trained per VFM (FeatUp (Fu et al.,

2024), JAFAR (Couairon et al., 2025), LoftUp (Huang et al., 2025)), and have recently advanced to methods that can be applied zero-shot to unseen VFMs without per-VFM training (AnyUp, NAF, RaysUp). These zero-shot methods, however, are trained with a single VFM as the teacher. When an upsampler is trained to reconstruct the features of several VFMs at once, it falls short of the best single-teacher model in NAF (Chambon et al., 2026, Tab. 10), and the multi-VFM version of AnyUp falls behind its single-VFM counterpart under the independent evaluation of DiveUp (Liu & Fan, 2026, Tab. 3). DiveUp attributes the degradation to a lack of upsampler capacity, and avoids it by keeping a single VFM as the reconstruction teacher and adding the consensus on spatial structure across several VFMs as an auxiliary teacher signal. AnyUp itself reports that multi-VFM training slightly lowers the score on DINOv2 relative to the model trained on DINOv2 alone (Wimmer et al., 2026, Tab. 7), and attributes this drop to the single-VFM model overfitting to the positional-encoding artifacts of DINOv2 (Wimmer et al., 2026, §5.2). It has not been examined which part of the upsampler design causes training with several reconstruction teachers to fail. This paper examines this question with a controlled comparison: all schemes use the VFM features as values and differ only in the positional encoding of the queries and keys.

Relative position in attention. There are two main ways to supply positional information to attention. The first adds position to the attention score as an independent term. The relative position representation of Shaw et al. (2018) adds the dot product of the query and a relative position vector, and stand-alone self-attention (Ramachandran et al., 2019) builds the relative position vector by concatenating embeddings of the row and column offsets. T5 and Swin Transformer add a learnable scalar for each displacement, and ALiBi (Press et al., 2022) adds a fixed bias proportional to distance. The second rotates queries and keys by an angle that depends on position, so that position multiplies content inside the dot product; RoPE is of this form. NAF uses the axial RoPE of RoPE-ViT (Heo et al., 2024) without modification. AnyUp and RayPE in RaysUp only adjust how the rotation angle is set or how the coordinates are taken, and none of the three has an independent term for position. Every existing cross-attention feature upsampler adopts the multiplicative form, and none has tested the additive form. In language models, TUPE (Ke et al., 2021) and DeBERTa (He et al., 2021) compare designs of the positional term within the additive form, while in ViTs, RoPE-ViT compares the additive form with the multiplicative form. All of these comparisons use the accuracy of one model trained on one objective as the criterion. This paper instead asks which form lets one upsampler learn from two teachers simultaneously.

Multi-teacher and multi-task learning. Prior work that distills multiple VFMs into a single model (Ranzinger et al., 2024; Heinrich et al., 2025; Chaybouti et al., 2026) treats the disagreement between teachers as a matter of feature distributions or loss weighting, and employs per-teacher heads. How the competing demands of the teachers interact within the shared backbone is left unexamined. In multi-task learning, gradient conflict, where two objectives produce opposing gradients on shared parameters, is a known cause of performance degradation. PCGrad projects conflicting gradients onto each other's normal planes, while CAGrad (Liu et al., 2021a) and Nash-MTL (Navon et al., 2022) alter how the update direction is computed. These methods, however, do not consistently outperform a simple sum of the losses (Kurin et al., 2022; Xin et al., 2022). The multi-teacher training of feature upsamplers employs no per-teacher heads, and its two reconstruction losses share every parameter. It is multi-task learning in its simplest form. This paper shows that the gradient conflict comes from the coupling form of the positional encoding, and uses PCGrad to test whether an optimization-side intervention can eliminate the degradation under multi-teacher training (§4.4).

3 Two Coupling Forms of Positional Encoding

3.1 Feature Upsampling and Attention Kernels

Among the feature upsamplers described in §2 that do not require per-VFM training, this paper builds on NAF, whose simple structure makes the effect of positional encoding easy to measure.

Formulation and neighborhood aggregation. Let the input high-resolution RGB image be $\mathbf{I} \in \mathbb{R}^{H_{\mathrm{HR}} \times W_{\mathrm{HR}} \times 3}$, and let the low-resolution feature map output by a frozen VFM be $\mathbf{F}^{\mathrm{LR}} \in \mathbb{R}^{H_{\mathrm{LR}} \times W_{\mathrm{LR}} \times D}$. Passing the input image through a lightweight convolutional image encoder Enc_θ gives a high-resolution feature map $\mathrm{Enc}_\theta(\mathbf{I})$. The image encoder Enc_θ consists of a pixel branch

with 1×1 convolutions and a context branch with 3×3 convolutions; §4.3 examines the gradients of these two layer types separately. We define the queries and keys as

$$Q_i = \mathrm{Enc}_\theta(\mathbf{I})_i, \qquad K_j = \frac{1}{|\mathcal{C}(j)|} \sum_{u \in \mathcal{C}(j)} \mathrm{Enc}_\theta(\mathbf{I})_u, \tag{1}$$

where $\mathcal{C}(j)$ is the set of high-resolution pixels contained in the low-resolution position j.

Following NAF, the keys to which a query at position i attends are restricted to a $K \times K$ window $\mathcal{N}(i)$ centered at the low-resolution position that contains i. The output feature $\widehat{\mathbf{F}}_i^{\mathrm{HR}}$ of the upsampler is the weighted average of the values $\mathbf{F}_j^{\mathrm{LR}}$ in the window with the attention weights $A(i,j)$,

$$\widehat{\mathbf{F}}_i^{\mathrm{HR}} = \sum_{j \in \mathcal{N}(i)} A(i,j) \mathbf{F}_j^{\mathrm{LR}}, \qquad A(i,j) = \mathrm{Softmax}_{j \in \mathcal{N}(i)}(S(i,j)) = \frac{\exp(S(i,j))}{\sum_{j' \in \mathcal{N}(i)} \exp(S(i,j'))}, \tag{2}$$

where $S(i,j)$ is the attention score before the softmax and $A(i,j)$ is the normalized attention weight.

Attention kernels and positional coupling. We call the set of attention weights $A(i,j)$ over the neighborhood window the **attention kernel**. Each key position j in the window is determined by its relative displacement $\Delta = j - i = (\Delta y, \Delta x)$ from the query, so the attention kernel can be viewed as a function of Δ. We call the model without positional encoding NONE and use it as the baseline for comparison in this paper (§4.1). The attention score of NONE is the inner product $S_{\text{None}}(i,j) = Q_i^\top K_j / \sqrt{d}$; it depends only on content similarity and not explicitly on Δ.

NAF (Chambon et al., 2026, §3) relates its attention, including RoPE, to the two components of classical joint bilateral filtering, and calls the inner product of queries and keys the content kernel and the RoPE factor the spatial kernel. We follow this naming: the part of the attention score $S(i,j)$ that gives content similarity is the **content kernel**, and the part that depends on the relative displacement Δ is the **spatial kernel**. The content kernel gives adaptivity: it aggregates from locations with similar content and prevents features from mixing across object boundaries. The spatial kernel gives locality: it aggregates preferentially from locations near the query position ($\Delta = 0$). Feature upsampling needs both kernels (Chambon et al., 2026; Wimmer et al., 2026).

Since S_{None} does not depend on Δ, only the positional encoding provides the spatial kernel, and the difference between positional encoding schemes appears as a difference in how the spatial kernel is combined with the content kernel. We define the way in which positional encoding is combined with the content kernel inside the score $S(i,j)$ as the **coupling form**, and compare two coupling forms in the next two subsections.

3.2 MULTIPLICATIVE COUPLING

RoPE rotates the queries and keys by position-dependent angles before the inner product (illustrated in the supplementary material). NAF applies RoPE to $\mathrm{Enc}_\theta(\mathbf{I})$ in equation 1 pixel by pixel before forming the queries and keys. The attention score is the inner product after rotation, and, by the composition rule of rotations, it depends on the relative displacement Δ rather than on absolute positions. This is the spatial kernel that RoPE provides.

To analyze the structure of this score, we decompose 2D RoPE in polar coordinates. RoPE splits the d channels into $d/2$ pairs c, assigns the vertical coordinate to the first half of the pairs and the horizontal coordinate to the second half, and gives each pair an angular frequency ω_c. Let Δ_c denote the coordinate difference along the axis that pair c handles. Writing the query and key of pair c in terms of amplitudes $r_{i,c}, r_{j,c}$ and phase angles $\alpha_{i,c}, \beta_{j,c}$, the score takes the form

$$S_{\text{RoPE}}(i,j) = \frac{1}{\sqrt{d}} \sum_{c=0}^{d/2-1} \underbrace{r_{i,c}\, r_{j,c}}_{\text{amplitude}} \cos\Big(\underbrace{\Psi_{i,j,c}}_{\text{content phase}} + \underbrace{\Delta\Phi_c(i,j)}_{\text{positional phase}} \Big), \tag{3}$$

where the **content phase difference** $\Psi_{i,j,c} = \beta_{j,c} - \alpha_{i,c}$ comes from the image features and the **positional phase difference** $\Delta\Phi_c(i,j) = \omega_c \Delta_c$ comes from the relative displacement (see the supplementary material for the derivation). Equation 3 shows that the amplitudes and the content

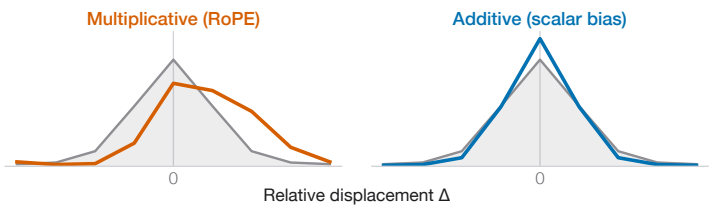


Figure 2: One-dimensional toy experiment. Attention distributions over the neighboring keys before (gray) and after (colored) a local content update centered at the query position. The pre-update distributions are identical for both schemes. **(a)** Multiplicative coupling (RoPE): the change in the content phase alone shifts the attention toward one side of the query and widens the distribution. **(b)** Additive coupling (SCALAR): the local decay centered at the query is largely preserved (details in the supplementary material).

phases are determined by the output of Enc_θ, and that position enters only as a phase shift inside the cosine. The same polar decomposition is derived in NAF (Chambon et al., 2026, Appendix A).

Expanding each term of equation 3 by the addition formula gives $r_{i,c}\, r_{j,c}(\cos\Psi_{i,j,c}\cos\Delta\Phi_c - \sin\Psi_{i,j,c}\sin\Delta\Phi_c)$, so the positional factors multiply the content factors (**multiplicative coupling**). The degrees of freedom to adjust the shape of the spatial kernel over the relative displacement Δ depend solely on the amplitude r and the content phase Ψ, both output by Enc_θ. NAF fixes the angular frequencies ω_c. Making them learnable still leaves position as a phase shift inside the cosine and does not remove this structural constraint (verified in §4.2).

To see how the requirement to adjust the spatial kernel reaches Enc_θ, we take the partial derivatives with respect to the amplitude r and the content phase Ψ, which gives

$$\begin{aligned}\frac{\partial S_{\text{RoPE}}(i,j)}{\partial \Psi_{i,j,c}} &= -\frac{1}{\sqrt{d}} r_{i,c}\, r_{j,c} \sin(\Psi_{i,j,c} + \Delta\Phi_c(i,j)),\\ \frac{\partial S_{\text{RoPE}}(i,j)}{\partial r_{i,c}} &= \frac{1}{\sqrt{d}} r_{j,c} \cos(\Psi_{i,j,c} + \Delta\Phi_c(i,j)).\end{aligned} \tag{4}$$

As shown in equation 4, both derivatives contain the positional phase $\Delta\Phi_c(i,j)$. The gradient of attention score reaches the queries and keys multiplied by a displacement-dependent coefficient, and the requirement to maintain the spatial kernel is propagated directly into the image encoder Enc_θ.

Hypothesis: One teacher's update disrupts the other's spatial kernel. Consider multi-teacher training, in which a single upsampler is trained with several VFMs as teachers (§2). Our hypothesis rests on two premises. First, a local decay centered at the query position ($\Delta = 0$) is needed regardless of which VFM serves as the teacher, but its extent and shape differ between teachers (§4.3). Second, the feature representation required of the queries and keys also differs between teachers. The features of DINOv3 have geometrically consistent local structure (Caron et al., 2021; Siméoni et al., 2026), whereas the features of SigLIP2 carry global semantics but contain positional misalignment and high-norm artifacts (Liu & Fan, 2026; Chaybouti et al., 2026).

To see what happens when the spatial kernel required by one teacher is formed by the content features, we use a one-dimensional toy experiment in Figure 2. Under multiplicative coupling (RoPE), a local content update alone changes the spatial kernel (Figure 2(a)). For a horizontal pair c, the cosine term $\cos(\Psi_{i,j,c} + \omega_c \Delta x)$ peaks at $\Delta x = -\Psi_{i,j,c}/\omega_c$, so when a feature update changes the content phase Ψ, the peak moves. When the update is local to the query position, $\Psi_{i,j,c}$ differs for each key and the shift differs for each Δx, so the spatial kernel shifts toward one side of the query and widens (Figure 2(a)). In the full model, the angular frequency ω_c differs across pairs, so the sum over all pairs flattens further. Under additive coupling (SCALAR), by contrast, the position bias determines the shape of the spatial kernel independently, so the decay centered at the query is largely preserved when the query features are updated (Figure 2(b); details in the supplementary material).

Our hypothesis is that the demand to keep the shape of the spatial kernel and the demand to acquire teacher-specific feature representations give opposing gradients to the shared encoder Enc_θ, producing gradient conflict in early training (verified in §4.3).

3.3 ADDITIVE COUPLING

In **additive coupling**, dedicated positional parameters that are independent of the content features determine the shape of the spatial kernel.

To verify that the gain of additive coupling does not depend on the expressiveness of the positional term, we compare two additive schemes with different expressiveness. **SCALAR** is the relative position bias of T5 and Swin Transformer, and adds a learnable scalar b (hereafter, the position bias) to the score for each relative displacement $(\Delta y, \Delta x)$. **VECTOR** is the axis-wise relative positional encoding of stand-alone self-attention, and adds the inner products of the learnable per-axis vectors $a_y(\Delta y), a_x(\Delta x) \in \mathbb{R}^{d/2}$ with the split query $Q_{i,y}, Q_{i,x}$.

$$S_{\text{Scalar}}(i,j) = \frac{Q_i^\top K_j}{\sqrt{d}} + b(\Delta y, \Delta x), \qquad S_{\text{Vector}}(i,j) = \frac{Q_i^\top K_j + Q_{i,y}^\top a_y(\Delta y) + Q_{i,x}^\top a_x(\Delta x)}{\sqrt{d}}. \tag{5}$$

SCALAR adds 1,156 parameters. In VECTOR, the positional term depends on the query but is added to the content similarity rather than multiplied by it. This term depends on Δ only through the learnable vectors a_y and a_x, so VECTOR is also an additive scheme.

Decoupling spatial constraints from the encoder. For contrast with multiplicative coupling (equation 4), the partial derivatives of the score of SCALAR, the representative additive coupling, with respect to the content phase difference $\Psi_{i,j,c}$ and the position bias b are

$$\frac{\partial S_{\text{Scalar}}(i,j)}{\partial \Psi_{i,j,c}} = -\frac{1}{\sqrt{d}} r_{i,c}\, r_{j,c} \sin \Psi_{i,j,c}, \qquad \frac{\partial S_{\text{Scalar}}(i,j)}{\partial b(\Delta y, \Delta x)} = 1. \tag{6}$$

Equation 6 differs from multiplicative coupling (equation 4) in two respects. First, the gradient with respect to the content phase difference contains no positional phase $\Delta\Phi$ and is identical to that in NONE, and so is the gradient with respect to the amplitude. Second, the gradient that modifies the shape of the spatial kernel reaches the position bias b with a coefficient of 1, and the update of b does not pass through the image encoder ($\partial b / \partial \theta = 0$; see the supplementary material). With this decoupling, the image encoder of SCALAR is not constrained to preserve spatial locality, and there is no pathway through which a feature update by one teacher changes the spatial kernel of the other (Figure 2(b)).

Testable predictions.

- **Prediction 1 (Order of the gain in multi-teacher transfer).** On tasks where fine spatial detail affects downstream performance, the gain from adding a second teacher is larger under additive coupling than under multiplicative coupling.
- **Prediction 2 (Gradient conflict in early training).** In early training, conflict between the two teachers' gradients, measured as negative cosine similarity over all image encoder parameters, occurs only under multiplicative coupling.
- **Prediction 3 (Limits of optimization-side intervention).** Suppose the performance drop under multiplicative coupling stems not from competing objectives but from the structural constraint that requires a single feature vector to both maintain the spatial kernel and describe content. Then an optimization-side intervention such as PCGrad, which projects conflicting gradients step by step, still leaves the gain of multiplicative coupling below that of additive coupling.

4 EXPERIMENTS

4.1 SETUP

Training. We use DINOv3 and SigLIP2 (Tschannen et al., 2025) as teachers. We train under three conditions: DINOv3 alone, SigLIP2 alone, and both teachers. Multi-teacher training shares all parameters including the positional encoding, and minimizes the equally weighted mean of the cosine-distance losses across the two teachers. We compare five schemes: NONE, RoPE (NAF), SCALAR, and VECTOR, as formulated in §3, as well as RoPE (learned freq.), whose angular frequencies ω_c are learnable. All models are trained on ImageNet-1k for 25,000 steps.

Table 1: Gain on each evaluation VFM for Cityscapes (mIoU) and DAVIS ($\mathcal{J}\&\mathcal{F}$). Under *Mean*, *Gain* averages the seven per-VFM gains, and *Score* denotes the absolute score of the multi-teacher run averaged over the same VFMs (best in bold, second best underlined). Columns group VFMs by training signal, and † marks the two teachers. The *Gain* column is shaded by coupling form: orange for multiplicative, blue for additive, and gray for None. *RoPE (NAF) + PCGrad* applies PCGrad at every training step (§4.4).

		Image only		Image–text				Distilled	Mean	
Scheme	Coupling	DINOv3†	FRANCA	CLIP	SigLIP2†	PE-Core	PE-Spatial	RADIO	Gain	Score
Cityscapes, mIoU										
None	–	+0.25	+0.72	+1.14	+1.27	+1.27	+0.81	+0.60	+0.87	55.62
RoPE (NAF)	mult.	−1.37	−1.11	−0.98	−0.19	−0.09	−0.78	−1.32	−0.83	58.57
RoPE (NAF) + PCGrad	mult.	−1.18	−0.97	−0.81	−0.07	+0.01	−0.66	−1.13	−0.68	58.72
RoPE (learned freq.)	mult.	−0.44	−0.23	−0.40	+0.38	+0.42	−0.09	−0.52	−0.13	59.24
Scalar	add.	−0.25	+0.30	+0.46	+1.10	+1.25	+0.51	+0.01	+0.48	**60.98**
Vector	add.	−0.08	+0.54	+0.85	+1.02	+0.97	+0.55	+0.18	+0.57	<u>60.79</u>
DAVIS, $\mathcal{J}\&\mathcal{F}$										
None	–	−1.80	−1.60	−1.90	−1.30	−0.70	−1.10	−1.80	−1.46	64.43
RoPE (NAF)	mult.	−3.90	−3.10	−2.80	−1.80	−2.50	−1.80	−3.70	−2.80	61.01
RoPE (NAF) + PCGrad	mult.	−3.30	−2.40	−2.20	−1.30	−1.40	−1.90	−2.40	−2.13	61.69
RoPE (learned freq.)	mult.	−3.80	−2.30	−2.20	−1.90	−1.90	−1.60	−3.20	−2.41	61.36
Scalar	add.	−2.60	−1.90	−1.60	−0.70	+0.10	−0.90	−2.40	−1.43	<u>64.64</u>
Vector	add.	−2.00	−1.60	−1.50	−0.30	+0.70	−0.70	−2.50	−1.13	**64.81**

Evaluation. Following NAF, we evaluate the features obtained by upsampling the output of a frozen VFM. For semantic segmentation (**Cityscapes** (Cordts et al., 2016) and **PASCAL VOC 2012**), we train a linear probe. For video object segmentation (**DAVIS 2017** (Pont-Tuset et al., 2017)), we use training-free label propagation (Caron et al., 2021). We evaluate on seven VFMs, consisting of the two teachers and five VFMs unseen during training (CLIP, FRANCA (Venkataramanan et al., 2026), PE-Core, PE-Spatial (Bolya et al., 2025), and RADIO), and report the mean score over the seven. The **gain** of a scheme is the score of its multi-teacher model (Multi) minus the higher score of its two single-teacher models (best Single).

4.2 Verification of Prediction 1: Multi-Teacher Transfer

Gain ordering across benchmarks. Table 1 shows the gain on each evaluation VFM. On Cityscapes, which requires precise boundaries for roads and pedestrians in driving scenes, additive coupling outperforms the best single teacher by $+0.48$ to $+0.57$ mIoU, whereas multiplicative coupling shows negative transfer of -0.13 to -0.83 mIoU. This contrast is consistent with Prediction 1 (§3.3). As the per-VFM entries of Table 1 show, the advantage of additive over multiplicative coupling holds on all seven VFMs. The gain of RoPE (NAF) is negative on all seven VFMs, and that of RoPE (learned freq.) is negative on five, all except SigLIP2 and PE-Core (mean -0.13 mIoU). Additive coupling attains a positive gain on six VFMs, all except DINOv3, which is one of the two teachers.

Unlike on Cityscapes, $\mathcal{J}\&\mathcal{F}$ dropped for all schemes on DAVIS, which tracks the pixels of an object across frames. NONE, which has no positional information, also dropped by 1.5%, indicating that a drop of this size occurs regardless of the coupling form. The drop under additive coupling stays around the level of NONE (1.1 to 1.4%), whereas the drop under multiplicative coupling (RoPE) reaches 2.8%, about twice that of NONE and additive coupling. Under multiplicative coupling, as the toy experiment in Figure 2 suggests, a feature update shifts and flattens the spatial kernel, and this geometric distortion accounts for the additional drop in video tracking, which requires pixel-accurate correspondence.

Table 2: Difference in multi-teacher scores relative to None, averaged over the seven evaluation VFMs (Cityscapes and VOC in mIoU, DAVIS in $\mathcal{J}\&\mathcal{F}$). Unlike Table 1, this table compares the multi-teacher scores across schemes. The DAVIS column, where the multiplicative schemes fall below None, is shaded by coupling form: orange for multiplicative, blue for additive.

Scheme	Cityscapes	VOC	DAVIS
RoPE (NAF)	+2.95	+0.62	−3.41
RoPE (learned freq.)	+3.62	+0.72	−3.07
Scalar	+5.36	+0.50	+0.21
Vector	+5.17	+0.41	+0.39

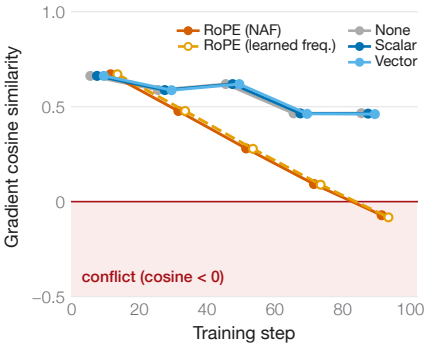


Figure 3: Cosine similarity between the DINOv3 and SigLIP2 gradients during the first 100 training steps, measured over all image encoder parameters. Only the two RoPE schemes decrease monotonically and turn negative (shaded region) in steps 80 to 99. The 100 steps are split into five bins of 20 steps; each point shows the mean within a bin. Points are offset horizontally by up to 4 steps for visibility; all schemes are measured in the same 20-step bins.

Table 3: Gradient alignment by parameter group over steps 50 to 99. The conflict under RoPE is localized in the 3×3 layers (shaded).

	Encoder 1×1		Encoder 3×3		Positional params	
Scheme	cos	confl.	cos	confl.	cos	confl.
None	$+0.39$	12%	$+0.54$	0%	–	–
RoPE (NAF)	$+0.31$	24%	-0.29	76%	–	–
RoPE (learned freq.)	$+0.30$	24%	-0.29	76%	-0.18	62%
Scalar	$+0.39$	12%	$+0.54$	0%	$+0.91$	0%
Vector	$+0.39$	12%	$+0.54$	0%	$+0.79$	0%

On PASCAL VOC, where recognizing coarse regions is the main requirement, the coupling forms hardly differ. The differences from NONE in Table 2 are $+0.41$ to $+0.72$ mIoU, nearly the same for all schemes and smaller than the $+2.95$ to $+5.36$ mIoU on Cityscapes. Coarse outlines suffice on VOC, so the presence of positional information itself has little effect on the score. The more a task requires fine, pixel-level spatial detail, the larger the difference between the coupling forms becomes, whereas on global tasks the difference remains small. This ordering across benchmarks is consistent with Prediction 1.

Ruling out model capacity. DiveUp attributes the performance drop in multi-teacher training to insufficient model capacity. However, model capacity does not explain the performance drop in our experiments. The angular frequencies of RoPE (NAF) are fixed, so RoPE (NAF) has exactly the same number of parameters as NONE. Yet NONE has a gain of $+0.87$ mIoU, whereas RoPE (NAF) has a gain of -0.83 mIoU. Among schemes with the same coupling form, adding parameters does not change the sign of the gain. RoPE (learned freq.) makes the angular frequencies learnable, but its gain remains negative. VECTOR has more parameters than SCALAR, and both gains are positive. Therefore, the performance drop under multiplicative coupling stems not from a shortage of model capacity but from the coupling form.

4.3 VERIFICATION OF PREDICTION 2: GRADIENT CONFLICT LOCALIZATION

To test Prediction 2, we analyze how the gradients of the two teachers interfere with each other in the image encoder. The positional parameters of additive coupling are initialized to zero, so an additive model is the same function as NONE at the start of training, and the effect of the coupling form on the gradients can be compared directly under the same initial conditions. Figure 3 shows the cosine similarity between the gradients of the two teachers, measured over all image encoder parameters, during the first 100 training steps. In the first 20 steps, all schemes show positive alignment ($\cos \approx +0.67$; the gradients of the two teachers point in the same direction). As training proceeds, the cosine decreases monotonically only for the two RoPE schemes and turns negative ($\cos < 0$) around step 80. Under NONE and additive coupling (SCALAR, VECTOR), by contrast, no conflict occurs during the first 100 steps. At the end of training, the cosine is negative under every scheme and is most negative under VECTOR, whose gain on Cityscapes is nonetheless positive (see the supplementary material). At a stationary point of the training objective, the two gradients sum to zero; such a point is Pareto stationary (Sener & Koltun, 2018). Unlike the early conflict, this late negative cosine is therefore not specific to multiplicative coupling.

To identify where this conflict occurs, we compute the cosine similarity between the gradients of the two teachers separately for the 1×1 and the 3×3 convolution layers of the image encoder (§3.1). We analyze steps 50 to 99, the latter half, in which the cosine of RoPE approaches negative values. The conflict rate is the fraction of steps with a negative cosine. Table 3 shows that the conflict occurs in the 3×3 layers of the two RoPE schemes. In the 1×1 layers, which have no spatial receptive field, all schemes maintain a positive mean cosine ($+0.30$ to $+0.39$) regardless of the coupling form. In the 3×3 layers, which have a spatial receptive field, NONE and additive coupling stay positive at every step with $\cos = +0.54$ (conflict rate 0%), whereas the two RoPE schemes turn negative with $\cos = -0.29$, reaching a conflict rate of 76%. That the conflict is concentrated in the 3×3 layers is consistent with the mechanism derived in §3. In the spatial convolution layers, the gradients conflict because a single feature representation must satisfy the different kernel shapes that the two teachers require. The position bias learned by SCALAR concentrates around the query after training with DINOv3 alone, and spreads along the vertical and horizontal axes after training with SigLIP2 alone (see the supplementary material).

Adding learnable parameters to the positional side does not remove the conflict under RoPE. In Table 3, the 3×3 layers of RoPE (learned freq.) with learnable frequencies show the same values as RoPE (NAF). The gradients of the learned frequencies also conflict between the two teachers at 62% of the steps. The gradients of the positional parameters of SCALAR and VECTOR, by contrast, have cosines of $+0.79$ to $+0.91$ and never conflict. Therefore, this conflict stems not from the absence of learnable parameters on the positional side but from multiplicative coupling.

4.4 VERIFICATION OF PREDICTION 3: LIMITS OF OPTIMIZATION-SIDE INTERVENTION

At steps where the inner product of the two teachers' gradients is negative, PCGrad projects each gradient onto the orthogonal complement of the other, which removes the conflicting component. As Figure 3 shows, under additive coupling and NONE, the conflict rate across all image encoder parameters is 0% in early training, and the PCGrad projection is never applied in the first 100 steps. By contrast, PCGrad applies 42 of its 200 candidate projections (two per step, one per teacher) for RoPE (NAF), whose 3×3 layers show a 76% conflict rate, and 48 for RoPE (learned freq.) (see the supplementary material). However, even when PCGrad is applied at every training step of RoPE (NAF), the gain on Cityscapes recovers only from -0.83 to -0.68 mIoU and remains negative (Table 1). The gap to the $+0.48$ to $+0.57$ mIoU gains of additive coupling is not closed.

The fact that the gain remains negative after PCGrad corrects the conflicting gradients is consistent with Prediction 3. PCGrad changes only the direction of the update. It does not change the structure of multiplicative coupling, in which the features of the image encoder must form the spatial kernels of both teachers at once (equations 3 and 4). Over steps 50 to 99, the cosine similarity across all image encoder parameters is $+0.05$ for RoPE (NAF) compared to $+0.50$ for additive coupling, indicating that the gradients of the two teachers are barely aligned on average (Figure 3). PCGrad only sets a negative inner product to zero, so the gradients remain barely aligned.

5 CONCLUSION AND LIMITATIONS

We show that the coupling form of positional encodings causes a performance drop when training a feature upsampler with multiple VFM teachers simultaneously. Under multiplicative coupling, which multiplies positional information into the features, a single feature must form a different spatial kernel for each teacher at once, leading to gradient conflict between the teachers. This conflict is not resolved by an optimization method that modifies the gradients, but by switching to additive coupling, which adds positional information to the attention score. An upsampler that learns from multiple VFMs should therefore add positional information to the score instead of multiplying it into the features.

Additive coupling turns the performance drop into a gain, but it remains unclear what determines the magnitude of this gain. One possible factor is that the upsampler uses the same spatial kernels for all teachers, even though different teachers require different kernels (§4.3). Adapting the kernels to each VFM may therefore make upsampling more effective, a direction we leave for future work.

Ethics statement

This work studies the positional encoding of feature upsamplers and involves no new data collection. All experiments use publicly available datasets (ImageNet-1k for training; Cityscapes, PASCAL VOC and DAVIS for evaluation) and publicly released vision foundation models, under their respective licenses. The upsamplers we train inherit whatever biases are present in those pretrained models and datasets; nothing in our method mitigates such biases, and the improvement we report is in feature quality, not in fairness or safety. We see no direct path from this work to harmful application beyond the risks already carried by the underlying foundation models.

AI use statement

During the preparation of this work, we used generative AI tools to assist with translating the manuscript into English. After using these tools, we reviewed and edited the content as needed and take full responsibility for the content of this paper.

Reproducibility statement

The training and evaluation setup is described in §4.1, and the full settings are listed in the supplementary material; all runs use seed 0 and are otherwise identical across positional schemes. The positional schemes are defined in §3 and §4.1, and their correspondence to the configuration names in the training code is given in the supplementary material. The gradient-observation and PCGrad protocols are described in §4.3 and §4.4. The supplementary material also includes the derivations behind §3, the full setup of the toy experiment, and the per-evaluation-VFM breakdown behind every averaged number. The code will be released upon acceptance.

References

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception Encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, pp. 67743–67796, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9630–9640, 2021.

Loick Chambon, Paul Couairon, Éloi Zablocki, Alexandre Boulch, Nicolas Thome, and Matthieu Cord. NAF: Zero-Shot Feature Upsampling via Neighborhood Attention Filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26604–26613, 2026.

Sofian Chaybouti, Sanath Narayan, Yasser Dahou, Phúc H. Lê Khac, Ankit Singh, Ngoc Dung Huynh, Wamiq Reyaz Para, Hilde Kuehne, and Hakim Hacid. SigLino: Efficient Multi-Teacher Distillation for Agglomerative Vision Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10071–10081, 2026.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

Paul Couairon, Loick Chambon, Louis Serrano, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. JAFAR: Jack up Any Feature at Any Resolution. In *Advances in Neural Information Processing Systems*, pp. 110050–110077, 2025.

Yuchuan Ding, Linfei Li, Lin Zhang, and Ying Shen. RaysUp: Ultra-light Universal Feature Upsampling via Geometry-Aware Ray Representation. In *European Conference on Computer Vision*, pp. 130–147, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

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 *International Conference on Learning Representations*, 2024.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*, 2021.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. RADIOv2.5: Improved Baselines for Agglomerative Vision Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22487–22497, 2025.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary Position Embedding for Vision Transformer. In *European Conference on Computer Vision*, pp. 289–305, 2024.

Haiwen Huang, Anpei Chen, Volodymyr Havrylov, Andreas Geiger, and Dan Zhang. LoftUp: Learning a Coordinate-Based Feature Upsampler for Vision Foundation Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9913–9923, 2025.

Guolin Ke, Di He, and Tie-Yan Liu. Rethinking Positional Encoding in Language Pre-training. In *International Conference on Learning Representations*, 2021.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In Defense of the Unitary Scalarization for Deep Multi-Task Learning. In *Advances in Neural Information Processing Systems*, pp. 12169–12183, 2022.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-Averse Gradient Descent for Multi-task Learning. In *Advances in Neural Information Processing Systems*, pp. 18878–18890, 2021a.

Wenze Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. Learning to Upsample by Learning to Sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6004–6014, 2023.

Xiaoqiong Liu and Heng Fan. DiveUp: Learning Feature Upsampling from Diverse Vision Foundation Models. arXiv preprint arXiv:2603.13571, March 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9992–10002, 2021b.

Hao Lu, Wenze Liu, Zixuan Ye, Hongtao Fu, Yuliang Liu, and Zhiguo Cao. SAPA: Similarity-Aware Point Affiliation for Feature Upsampling. In *Advances in Neural Information Processing Systems*, pp. 20889–20901, 2022.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-Task Learning as a Bargaining Game. In *International Conference on Machine Learning*, pp. 16428–16446, 2022.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. arXiv preprint arXiv:1704.00675, April 2017.

Ofir Press, Noah A. Smith, and Mike Lewis. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *International Conference on Learning Representations*, 2022.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-Alone Self-Attention in Vision Models. In *Advances in Neural Information Processing Systems*, 2019.

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.

Ozan Sener and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. In *Advances in Neural Information Processing Systems*, 2018.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-Attention with Relative Position Representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 464–468, 2018.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv preprint arXiv:2502.14786, February 2025.

Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Elias Ramzi, Spyros Gidaris, Andrei Bursuc, and Yuki M. Asano. Franca: Nested Matryoshka Clustering for Scalable Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10533–10544, 2026.

Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. CARAFE: Content-Aware ReAssembly of FEatures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3007–3016, 2019.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. AnyUp: Universal Feature Upsampling. In *International Conference on Learning Representations*, 2026.

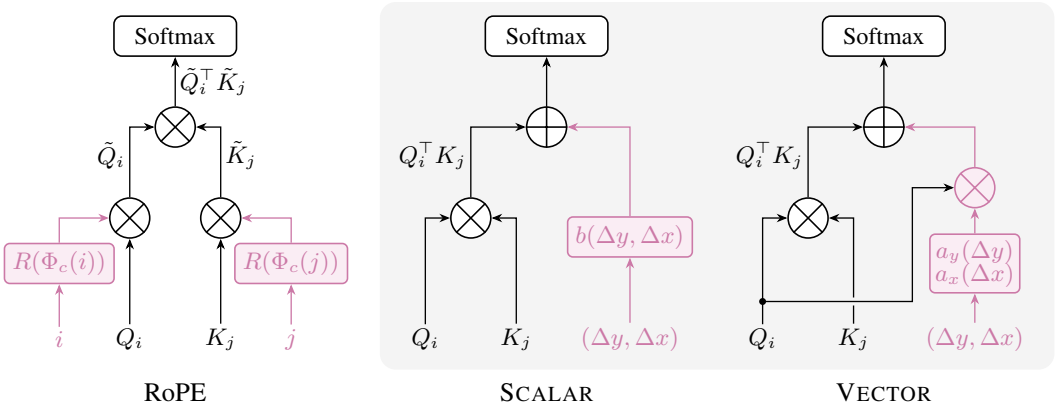
Derrick Xin, Behrooz Ghorbani, Ankush Garg, Orhan Firat, and Justin Gilmer. Do Current Multi-Task Optimization Methods in Deep Learning Even Help? In *Advances in Neural Information Processing Systems*, pp. 13597–13609, 2022.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient Surgery for Multi-Task Learning. In *Advances in Neural Information Processing Systems*, pp. 5824–5836, 2020.

SUPPLEMENTARY MATERIAL

A DERIVATION OF THE ATTENTION SCORE UNDER ROPE AND THE GRADIENTS UNDER THE TWO COUPLING FORMS

Equation 3 in the main text expresses the attention score under RoPE in terms of the cosine of the sum of the content and positional phase differences. Section 3.3 states that under SCALAR, updates to the position bias do not pass through the image encoder. This section derives equation 3 from the definition of 2D RoPE in NAF and compares the gradients with respect to the image encoder under the two coupling forms. Figure S1 shows a schematic of the two coupling forms.



(a) Multiplicative **(b) Additive**

Figure S1: Two coupling forms of positional encoding. **(a)** Multiplicative coupling (RoPE) rotates Q_i and K_j by angles determined by their positions i and j before the inner product, so position is multiplied into the content features. Only the output of the shared image encoder can change the shape of the spatial kernel. **(b)** Additive coupling adds a separate positional term to the content similarity, and dedicated parameters (b in SCALAR; a_y and a_x in VECTOR) determine the shape of the spatial kernel.

Notation. We follow the notation of §3.2. Let $Q_i \in \mathbb{R}^d$ be the content vector at query position $i = (i_y, i_x)$, and let $K_j \in \mathbb{R}^d$ be the content vector at key position $j = (j_y, j_x)$. The number of channels per head, d, is divisible by four. In this section we write Ψ_c for $\Psi_{i,j,c}$, $\Delta\Phi_c$ for $\Delta\Phi_c(i,j)$, and $S(i,j)$ for the score under either coupling form.

Attention score under RoPE. RoPE groups the d channels into $d/2$ pairs and treats the c-th pair as a two-dimensional vector,

$$Q_{i,c} = \begin{pmatrix} Q_{i,2c} \\ Q_{i,2c+1} \end{pmatrix}, \qquad K_{j,c} = \begin{pmatrix} K_{j,2c} \\ K_{j,2c+1} \end{pmatrix}, \qquad c = 0, \ldots, \frac{d}{2} - 1. \tag{S1}$$

Each pair has an angular frequency ω_c. The rotation angle at position i uses the vertical coordinate for the first half of the pairs and the horizontal coordinate for the second half, with the coordinates normalized to $[-1, 1]$ along each axis,

$$\Phi_c(i) = \begin{cases} \omega_c \, i_y, & 0 \le c < d/4, \\ \omega_c \, i_x, & d/4 \le c < d/2. \end{cases} \tag{S2}$$

Let $R(\Phi)$ be the two-dimensional rotation matrix for angle Φ. The rotated pairs are $\tilde{Q}_{i,c} = R(\Phi_c(i))\, Q_{i,c}$ and $\tilde{K}_{j,c} = R(\Phi_c(j))\, K_{j,c}$, and the attention score is

$$S(i,j) = \frac{1}{\sqrt{d}} \sum_{c=0}^{d/2-1} \tilde{Q}_{i,c}^\top \tilde{K}_{j,c} = \frac{1}{\sqrt{d}} \sum_{c=0}^{d/2-1} Q_{i,c}^\top \, R(\Phi_c(i))^\top R(\Phi_c(j)) \, K_{j,c}. \tag{S3}$$

NAF rotates the features and then average-pools them into keys. We write the rotation angle of a high-resolution pixel u within the low-resolution position j as the angle $\Phi_c(j)$ at the center of j plus a fixed offset $\delta_{u,c}$. Since $R(\Phi_c(j) + \delta_{u,c}) = R(\Phi_c(j))\, R(\delta_{u,c})$ and averaging is linear,

$$\tilde{K}_{j,c} = R(\Phi_c(j))\, \bar{K}_{j,c}, \qquad \bar{K}_{j,c} = \frac{1}{|\mathcal{C}(j)|} \sum_{u \in \mathcal{C}(j)} R(\delta_{u,c})\, K_{u,c}, \tag{S4}$$

where $\mathcal{C}(j)$ is the set of pixels inside j. The amplitude $r_{j,c}$ and the phase $\beta_{j,c}$ in equation 3 are those of $\bar{K}_{j,c}$. Below we write $K_{j,c}$ for $\bar{K}_{j,c}$.

Rotation matrices satisfy $R(\Phi)^\top = R(-\Phi)$, so $R(\Phi_c(i))^\top R(\Phi_c(j)) = R(\Phi_c(j) - \Phi_c(i))$. With the positional phase difference

$$\Delta\Phi_c = \Phi_c(j) - \Phi_c(i) = \begin{cases} \omega_c\,(j_y - i_y), & 0 \le c < d/4, \\ \omega_c\,(j_x - i_x), & d/4 \le c < d/2, \end{cases} \tag{S5}$$

the score becomes

$$S(i,j) = \frac{1}{\sqrt{d}} \sum_{c=0}^{d/2-1} Q_{i,c}^\top\, R(\Delta\Phi_c)\, K_{j,c}. \tag{S6}$$

The rotation in the score depends only on the coordinate difference between the two positions.

We write each pair in polar coordinates,

$$Q_{i,c} = r_{i,c} \begin{pmatrix} \cos\alpha_{i,c} \\ \sin\alpha_{i,c} \end{pmatrix}, \qquad K_{j,c} = r_{j,c} \begin{pmatrix} \cos\beta_{j,c} \\ \sin\beta_{j,c} \end{pmatrix}. \tag{S7}$$

A rotation by $\Delta\Phi_c$ adds $\Delta\Phi_c$ to the angle, so $R(\Delta\Phi_c)\, K_{j,c} = r_{j,c}\,(\cos(\beta_{j,c} + \Delta\Phi_c),\ \sin(\beta_{j,c} + \Delta\Phi_c))^\top$. Taking the inner product with $Q_{i,c}$ and applying the cosine addition formula gives

$$Q_{i,c}^\top R(\Delta\Phi_c)\, K_{j,c} = r_{i,c}\, r_{j,c} \cos(\beta_{j,c} - \alpha_{i,c} + \Delta\Phi_c). \tag{S8}$$

Setting $\Psi_c = \beta_{j,c} - \alpha_{i,c}$ and summing over the pairs gives equation 3,

$$S(i,j) = \frac{1}{\sqrt{d}} \sum_{c=0}^{d/2-1} r_{i,c}\, r_{j,c} \cos(\Psi_c + \Delta\Phi_c). \tag{S9}$$

When $\Delta\Phi_c = 0$, the rotation between the query and the key is the identity and the right-hand side equals $Q_i^\top K_j / \sqrt{d}$. The same polar decomposition is derived in NAF (Chambon et al., 2026, Appendix A), which interprets it as the inverse discrete Fourier transform of the aggregation kernel, whereas this paper uses the decomposition to track which parameters the gradient reaches.

Gradients under RoPE. Equation 4 in the main text gives the derivatives of equation 3 with respect to the content phase and the amplitude. Both coefficients depend on $\Delta\Phi_c$. By the chain rule, the gradient of the loss $\mathcal{L}$ with respect to the image encoder parameters θ is

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i,j} \frac{\partial \mathcal{L}}{\partial S(i,j)} \sum_c \left[\frac{\partial S(i,j)}{\partial \Psi_c} \frac{\partial \Psi_c}{\partial \theta} + \frac{\partial S(i,j)}{\partial r_{i,c}} \frac{\partial r_{i,c}}{\partial \theta} + \frac{\partial S(i,j)}{\partial r_{j,c}} \frac{\partial r_{j,c}}{\partial \theta} \right]. \tag{S10}$$

When a teacher requires a different shape of the spatial kernel, the requirement appears as $\partial\mathcal{L}/\partial S(i,j)$. In the gradient with respect to θ, it is multiplied by coefficients that depend on $\Delta\Phi_c$. RoPE has no other learnable parameters that change the shape of the spatial kernel, so a requirement on the kernel shape can be met only by changing the image encoder parameters θ. Making the frequencies ω_c learnable, as in RoPE (learned freq.), only changes the scale of $\Delta\Phi_c$. The equations above keep the same form. The gradient with respect to a learned frequency, with the pooled key $K_{j,c}$ held fixed, is

$$\frac{\partial S(i,j)}{\partial \omega_c} = -\frac{r_{i,c}\, r_{j,c}}{\sqrt{d}}\, \Delta_c \sin(\Psi_c + \Delta\Phi_c), \tag{S11}$$

where Δ_c is the coordinate difference along the axis of pair c. This coefficient contains the content phase Ψ_c, so the update of a frequency also depends on the content features.

Gradients under additive coupling. For SCALAR equation 5, $S(i,j) = Q_i^\top K_j/\sqrt{d} + b(\Delta)$, so

$$\frac{\partial S(i,j)}{\partial b(\Delta)} = 1, \qquad \frac{\partial b(\Delta)}{\partial \theta} = 0, \qquad \frac{\partial S(i,j)}{\partial \theta} = \frac{1}{\sqrt{d}} \frac{\partial \left(Q_i^\top K_j\right)}{\partial \theta}. \tag{S12}$$

The derivative of the score with respect to $b(\Delta)$ is one, so a requirement on the kernel shape changes the position bias directly. The gradient of the score with respect to θ contains only the content similarity $Q_i^\top K_j$, so in the gradient of the loss with respect to θ, the coefficient that multiplies $\partial \mathcal{L}/\partial S(i,j)$ does not depend explicitly on Δ, as in NONE. The positional term in VECTOR, $(Q_{i,y}^\top a_y(\Delta y) + Q_{i,x}^\top a_x(\Delta x))/\sqrt{d}$, which we write as $Q_i^\top a(\Delta)/\sqrt{d}$, depends on Q_i, so the gradient with respect to θ that passes through Q_i contains $a(\Delta)$. As in SCALAR, this term is added to the content similarity rather than multiplied by it. Since $\partial S(i,j)/\partial a(\Delta) = Q_i/\sqrt{d}$, a requirement on the kernel shape reaches the parameters $a(\Delta)$ directly, just as it reaches $b(\Delta)$ under SCALAR.

B THE ONE-DIMENSIONAL TOY EXPERIMENT

Figure 2 in §3.2 uses a one-dimensional example to show that updating the content features changes the attention distribution under RoPE but largely preserves it under SCALAR. This section describes the setup.

Setup. The toy model has positions $r = -13, \ldots, 13$ on a line, with the query at $r = 0$. Each position has a two-dimensional unit vector

$$g_\theta(r) = (\cos(\theta a_r),\ \sin(\theta a_r))^\top, \qquad a_r = \exp(-r^2/18). \tag{S13}$$

This feature corresponds to the output of the image encoder. A single learnable scalar θ, initialized to 0, rotates the features, and the rotation is largest at the query ($a_0 = 1$). This rotation represents a local content update at the query position. As in NAF, both the query and the keys are built from the features. The query is $Q = h_0$, and each of the nine keys $\Delta = -4, \ldots, 4$ is the average of h_r over three adjacent positions,

$$K_\Delta = \frac{1}{3} \sum_{r=3\Delta-1}^{3\Delta+1} h_r, \qquad h_r = \begin{cases} R(\omega r/3)\, g_\theta(r) & \text{(RoPE)}, \\ g_\theta(r) & \text{(SCALAR)}, \end{cases} \tag{S14}$$

where $R(\phi)$ is a rotation by angle ϕ and $\omega = \pi/4$. The coordinate $r/3$ places key Δ at Δ. The attention p_θ is the softmax of the scores S_Δ over the nine keys, where

$$S_\Delta = \beta\, Q^\top K_\Delta \quad \text{(RoPE)}, \qquad S_\Delta = \beta\, Q^\top K_\Delta + b_\Delta \quad \text{(SCALAR)}, \tag{S15}$$

and $\beta = 2$. The fixed position bias $b_\Delta = \beta\{\kappa \cos(\omega\Delta) - 1\}$, where $\kappa = (1 + 2\cos(\omega/3))/3$, makes the scores of the two schemes identical at $\theta = 0$.

Content update. We minimize $1 - \cos(\theta - \theta^*)$, which drives the feature at the query toward $\theta^* = 60°$, using Adam with a learning rate of 0.05 for 200 steps. Both schemes reach $\theta \approx 60°$. The change in attention, measured by $\mathrm{KL}(p_0 \,\|\, p_\theta)$ from the distribution p_0 at $\theta = 0$, is 0.263 for RoPE and 0.033 for SCALAR. Under SCALAR, the position bias b_Δ is fixed, so this remaining change comes solely from the content term $\beta\, Q^\top K_\Delta$.

C EXPERIMENTAL DETAILS

Section 4.1 states only the main settings of training and evaluation. This section provides the complete settings and the correspondence between the scheme names and the configuration names in the training code.

Training and evaluation. Table S1 summarizes the training and evaluation settings, and Table S2 lists the seven evaluation VFMs, all of which are Base-size models. The architecture and the linear probe follow the official NAF implementation. Label propagation upsamples the features by a factor of two, as in NAF, and uses the full frame as guidance.

Table S1: Training and evaluation settings.

Setting	Value
Training	
Data	ImageNet-1k
Image size	512×512 (target), 256×256 (input features), 128×128 (guidance)
Teachers	DINOv3, SigLIP2, or both
Loss	cosine distance, averaged over teachers
Batch size	4
Optimizer	AdamW
Learning rate	2×10^{-4}, constant
Weight decay	10^{-5}
Steps	25,000
Architecture	NAF (256 channels, 4 heads, 9×9 window)
Linear probe (Cityscapes, VOC)	
Image size	448×448, center crop
Optimizer	AdamW
Learning rate	5×10^{-4}, cosine annealing
Epochs	20
Label propagation (DAVIS)	
Frames	480p
Context	first frame and 7 preceding frames
Label resolution	$2\times$ VFM feature map
Propagation window	± 12 positions on the label grid
Top-k	5

Table S2: The seven evaluation VFMs, all Base-size models, in the column order of Table 1. DINOv3 and SigLIP2 also serve as teachers. RADIO and FRANCA are loaded via torch.hub from their official repositories, and the other five from timm.

VFM	Checkpoint	Patch size
DINOv3 (Siméoni et al., 2026)	`vit_base_patch16_dinov3.lvd1689m`	16
FRANCA (Venkataramanan et al., 2026)	`franca_vitb14`	14
CLIP (Radford et al., 2021)	`vit_base_patch16_clip_384`	16
SigLIP2 (Tschannen et al., 2025)	`vit_base_patch16_siglip_512.v2_webli`	16
PE-Core (Bolya et al., 2025)	`vit_pe_core_base_patch16_224.fb`	16
PE-Spatial (Bolya et al., 2025)	`vit_pe_spatial_base_patch16_512.fb`	16
RADIO (Heinrich et al., 2025)	`radio_v2.5-b`	16

Schemes. RoPE (NAF) is the official NAF configuration (Table S3), applied per attention head. It uses 16 fixed angular frequencies $\omega_t = 2\pi/100^{t/16}$, $t = 0, \ldots, 15$, shared by the vertical and horizontal pairs, and normalizes the coordinates to $[-1, 1]$ along each axis. RoPE (learned freq.) multiplies each ω_t by $\exp(\gamma_t)$, where the 16 parameters γ_t are learned, initialized to zero, and excluded from weight decay. In a 9×9 window that is shifted at the image border, the displacement along each axis ranges from -8 to 8. SCALAR has one scalar per head and displacement ($4 \times 17^2 = 1{,}156$ parameters), and VECTOR has one 32-dimensional vector per head, axis, and displacement along that axis ($4 \times 2 \times 17 \times 32 = 4{,}352$ parameters). Because the positional parameters of SCALAR and VECTOR are initialized to zero, every model with additive coupling is the same function as NONE at step 0.

D RESULTS ON EACH EVALUATION VFM

Table S4 splits the gain in Table 1 into the mean over the two teachers and the mean over the five unseen VFMs. Tables S5, S6, and S7 show the per-VFM scores on Cityscapes, VOC, and DAVIS for the single-teacher and multi-teacher models.

Note that the gain in Table 1 is computed by taking, for each VFM, the difference between the Multi score and the higher of the two single-teacher scores, and then averaging these differences. The difference between the Mean columns instead compares Multi with the same single-teacher model

Table S3: Scheme names in this paper, the configuration names in the training code, and the positional parameters of each scheme. In the Coupling column, "mult." and "add." denote multiplicative and additive coupling, respectively.

Scheme	Configuration name	Coupling	Positional parameters	% of image encoder
NONE	`none`	–	none	–
RoPE (NAF)	`rope_official`	mult.	none	–
RoPE (learned freq.)	`rope_naf_learnable`	mult.	16	< 0.01
SCALAR	`learned_rpb`	add.	1,156	0.17
VECTOR	`sasa`	add.	4,352	0.66

on all seven VFMs, so the two values differ when the better single-teacher model varies by VFM, as with NONE, SCALAR, and VECTOR on Cityscapes.

Table S4: Gain on Cityscapes (mIoU) and DAVIS ($\mathcal{J}\&\mathcal{F}$), averaged over all seven evaluation VFMs, over the two teachers (DINOv3 and SigLIP2), and over the five VFMs unseen during training. In the Coupling column, "mult." and "add." denote multiplicative and additive coupling, respectively.

		Cityscapes mIoU			DAVIS $\mathcal{J}\&\mathcal{F}$		
Scheme	Coupling	All 7	Teachers	Unseen 5	All 7	Teachers	Unseen 5
None	–	$+0.87$	$+0.76$	$+0.91$	-1.46	-1.55	-1.42
RoPE (NAF)	mult.	-0.83	-0.78	-0.86	-2.80	-2.85	-2.78
RoPE (learned freq.)	mult.	-0.13	-0.03	-0.17	-2.41	-2.85	-2.24
Scalar	add.	$+0.48$	$+0.42$	$+0.51$	-1.43	-1.65	-1.34
Vector	add.	$+0.57$	$+0.47$	$+0.61$	-1.13	-1.15	-1.12

Table S5: Evaluation scores on Cityscapes (mIoU) for each VFM. DINOv3 and SigLIP2 are the teachers. "Single DINOv3" and "Single SigLIP2" are trained with a single teacher, while "Multi" uses both. "Multi+PCGrad" applies PCGrad at every training step (§4.4).

Scheme	Training	DINOv3	FRANCA	CLIP	SigLIP2	PE-Core	PE-Spatial	RADIO	Mean
None	Single DINOv3	59.39	58.06	51.19	51.81	50.44	54.57	56.40	54.55
None	Single SigLIP2	56.89	56.57	50.96	52.63	51.04	53.33	54.69	53.73
None	Multi	59.64	58.79	52.32	53.91	52.30	55.39	57.00	55.62
RoPE (NAF)	Single DINOv3	63.91	62.04	56.13	57.91	55.24	59.63	60.98	59.41
RoPE (NAF)	Single SigLIP2	59.98	58.75	52.88	55.59	53.11	56.61	57.21	56.31
RoPE (NAF)	Multi	62.54	60.93	55.14	57.72	55.15	58.86	59.66	58.57
RoPE (NAF)	Multi+PCGrad	62.73	61.07	55.32	57.84	55.26	58.97	59.85	58.72
RoPE (learned freq.)	Single DINOv3	63.86	62.00	56.09	57.87	55.20	59.63	60.94	59.37
RoPE (learned freq.)	Single SigLIP2	62.41	60.95	54.45	57.05	54.21	58.52	59.27	58.12
RoPE (learned freq.)	Multi	63.42	61.77	55.68	58.25	55.61	59.54	60.42	59.24
Scalar	Single DINOv3	66.75	63.85	56.56	57.40	54.24	61.15	62.76	60.39
Scalar	Single SigLIP2	64.63	62.46	55.54	57.75	54.69	59.93	60.87	59.41
Scalar	Multi	66.50	64.15	57.02	58.85	55.93	61.67	62.77	60.98
Vector	Single DINOv3	66.42	63.48	56.05	56.81	53.48	60.86	62.44	59.94
Vector	Single SigLIP2	64.51	62.41	55.44	57.60	54.69	59.78	60.82	59.32
Vector	Multi	66.34	64.01	56.90	58.62	55.65	61.41	62.62	60.79

Section 4.2 states that model capacity does not cause negative transfer, because the gain of RoPE (learned freq.) is also negative. The SigLIP2 column of Table S5 shows that Single SigLIP2 outperforms Single DINOv3 under NONE and additive coupling, but falls 2.3 mIoU below it under RoPE (NAF). The Mean column also shows that Single SigLIP2 under RoPE (learned freq.) scores 1.8 mIoU higher than under RoPE (NAF), and that its gap to Single DINOv3 approaches the gaps under NONE and additive coupling. Therefore, RoPE (learned freq.) trained with both teachers falls below Single DINOv3 not because it cannot learn from SigLIP2, but because it learns from the two teachers at once. This is consistent with the gradient conflict in the 3×3 layers observed in §4.3.

Table S6: Evaluation scores on PASCAL VOC (mIoU) for each VFM. DINOv3 and SigLIP2 are the teachers. "Single DINOv3" and "Single SigLIP2" are trained with a single teacher, while "Multi" uses both.

Scheme	Training	DINOv3	FRANCA	CLIP	SigLIP2	PE-Core	PE-Spatial	RADIO	Mean
None	Single DINOv3	85.96	82.99	78.93	77.33	73.21	75.85	85.06	79.90
None	Single SigLIP2	84.51	82.63	79.55	79.70	76.15	75.36	84.42	80.33
None	Multi	85.93	83.47	80.12	79.66	75.77	76.29	85.52	80.97
RoPE (NAF)	Single DINOv3	86.93	84.08	80.94	79.47	74.77	77.14	86.64	81.42
RoPE (NAF)	Single SigLIP2	85.16	82.59	79.72	79.81	75.38	75.85	84.81	80.47
RoPE (NAF)	Multi	86.38	83.75	81.05	80.61	76.16	76.94	86.21	81.58
RoPE (learned freq.)	Single DINOv3	86.94	84.08	80.94	79.47	74.78	77.15	86.65	81.43
RoPE (learned freq.)	Single SigLIP2	85.82	83.17	80.05	79.69	75.15	76.40	85.33	80.80
RoPE (learned freq.)	Multi	86.59	83.96	81.14	80.56	76.10	77.10	86.36	81.69
Scalar	Single DINOv3	87.98	84.20	79.75	76.09	70.91	77.48	86.53	80.42
Scalar	Single SigLIP2	86.84	83.90	80.44	79.77	75.16	77.00	86.07	81.31
Scalar	Multi	87.74	84.41	80.71	79.00	74.03	77.63	86.71	81.46
Vector	Single DINOv3	87.85	84.03	79.38	75.34	70.17	77.31	86.35	80.06
Vector	Single SigLIP2	86.78	83.87	80.47	79.56	75.12	77.03	86.06	81.27
Vector	Multi	87.67	84.38	80.67	78.74	73.81	77.64	86.70	81.37

Table S7: Evaluation scores on DAVIS ($\mathcal{J}\&\mathcal{F}$, %) for each VFM. DINOv3 and SigLIP2 are the teachers. "Single DINOv3" and "Single SigLIP2" are trained with a single teacher, while "Multi" uses both. "Multi+PCGrad" applies PCGrad at every training step (§4.4).

Scheme	Training	DINOv3	FRANCA	CLIP	SigLIP2	PE-Core	PE-Spatial	RADIO	Mean
None	Single DINOv3	69.10	72.10	65.40	65.00	49.50	70.30	69.80	65.89
None	Single SigLIP2	64.30	66.80	60.00	61.50	47.90	66.10	65.60	61.74
None	Multi	67.30	70.50	63.50	63.70	48.80	69.20	68.00	64.43
RoPE (NAF)	Single DINOv3	65.30	68.80	62.80	64.50	48.40	69.20	67.70	63.81
RoPE (NAF)	Single SigLIP2	58.40	63.20	57.60	59.10	42.20	63.70	60.60	57.83
RoPE (NAF)	Multi	61.40	65.70	60.00	62.70	45.90	67.40	64.00	61.01
RoPE (NAF)	Multi+PCGrad	62.00	66.40	60.60	63.20	47.00	67.30	65.30	61.69
RoPE (learned freq.)	Single DINOv3	65.50	68.80	62.90	64.20	48.40	68.90	67.70	63.77
RoPE (learned freq.)	Single SigLIP2	59.40	63.60	58.00	58.80	43.00	64.60	60.50	58.27
RoPE (learned freq.)	Multi	61.70	66.50	60.70	62.30	46.50	67.30	64.50	61.36
Scalar	Single DINOv3	70.40	72.90	66.70	64.50	46.70	71.70	69.60	66.07
Scalar	Single SigLIP2	62.70	67.40	62.00	60.30	46.00	67.10	63.80	61.33
Scalar	Multi	67.80	71.00	65.10	63.80	46.80	70.80	67.20	64.64
Vector	Single DINOv3	70.10	72.60	66.90	64.50	46.60	71.40	69.50	65.94
Vector	Single SigLIP2	63.00	67.50	62.60	61.40	45.60	67.50	63.30	61.56
Vector	Multi	68.10	71.00	65.40	64.20	47.30	70.70	67.00	64.81

E POSITION BIAS LEARNED UNDER EACH TEACHER

Section 4.3 attributes the gradient conflict under RoPE to the different spatial kernels that the two teachers require. Under SCALAR, the positional term of the score is the position bias $b(\Delta y, \Delta x)$, an image-independent learned parameter, so the position bias learned with a single teacher reflects the kernel shape that this teacher requires. This section compares the position biases learned with DINOv3 alone, with SigLIP2 alone, and with both teachers (Figure S2 and Table S8), and discusses what the comparison implies for the multi-teacher model.

The three runs initialize the position bias to zero and differ only in their teachers. A constant added to the position bias of a head does not change the softmax, so the figure and table use the position bias with the head-wise mean subtracted.

SCALAR learns a different position bias with each teacher. Taken alone, the position bias learned with DINOv3 places 98% of the attention weight within the 3×3 neighborhood of the query, whereas the position bias learned with SigLIP2 places only 49% there and spreads its weight along the vertical and horizontal axes (Figure S2). These two position biases are weakly correlated. The position

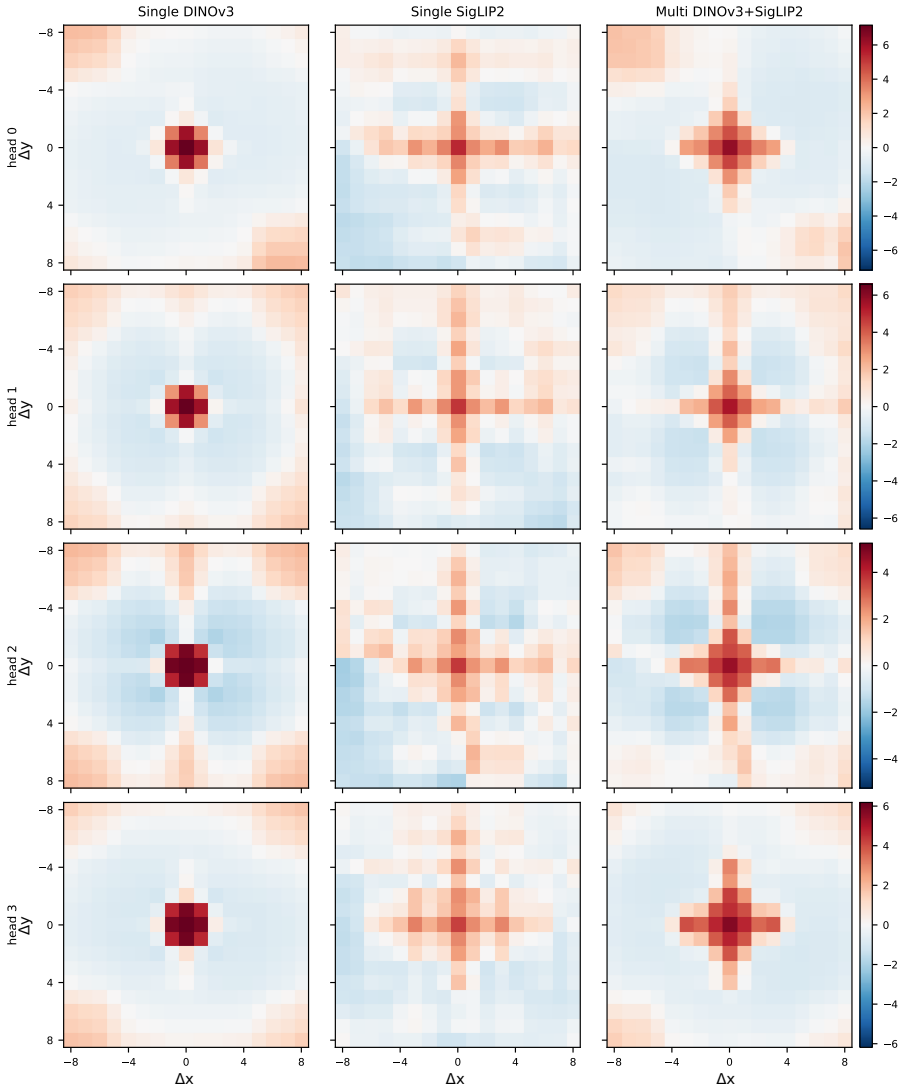


Figure S2: The position bias $b(\Delta y, \Delta x)$ of SCALAR after 25,000 steps of training with DINOv3 alone, SigLIP2 alone, and both teachers. Each row corresponds to one of the four heads. Color shows the difference from the mean of the head, and the scale is shared within each row. The values concentrate around the query for DINOv3, spread along the vertical and horizontal axes for SigLIP2, and take an intermediate shape for both teachers.

Table S8: The position bias of SCALAR at 25,000 steps under the three training conditions (mean over the four heads). The attention weight (%) is computed from the position bias alone over the 9×9 window. Pearson is the correlation between the mean-subtracted 17×17 position biases.

	Single DINOv3	Single SigLIP2	Multi
Attention weight within $\lvert\Delta y\rvert, \lvert\Delta x\rvert \leq 1$ (%)	98.0	49.2	72.9
Pearson correlation with Single DINOv3	–	0.311	0.773
Pearson correlation with Single SigLIP2	0.311	–	0.699

bias learned with both teachers places an intermediate proportion within the 3×3 neighborhood and is closer to each single-teacher position bias than the two single-teacher position biases are to each other (Table S8).

The upsampler reconstructs the features of both teachers using the same position bias, and the position bias learned with both teachers lies between the two single-teacher position biases (Table S8). At inference, the position bias is fixed and the queries and keys are computed from the image alone (equation 1 in §3.1), so the upsampler applies this intermediate spatial kernel to every VFM, including the five VFMs unseen during training. Under SCALAR, the compromise between the spatial kernels required by the two teachers is thus made in the position bias, which the image encoder does not produce (§3.3). Under RoPE, the same compromise must be made in the features of the image encoder, where the gradients of the two teachers conflict early in training (§4.3). The intermediate position bias still gives neither teacher the kernel it requires, and §5 identifies this shared kernel as a possible limit on the gain.

F GRADIENT MEASUREMENTS AND PCGRAD PROJECTIONS

Section 4.3 shows that under RoPE, the gradients of the two teachers conflict early in training, with the conflict localized in the 3×3 layers of the image encoder, and that the cosine similarity is negative under every scheme at the end of training. Section 4.4 shows that PCGrad does not close the gap. This section describes the measurement setup, shows where the negative cosine similarity at the end of training lies, and relates the PCGrad projections to the early conflict.

Measurement. Gradients during the first 100 steps are measured in the multi-teacher training run of each scheme. At each step, the loss for each teacher is backpropagated separately on the same batch before the update. The cosine similarity in Figure 3 of the main text is computed over the $662{,}528$ parameters of the image encoder, and the gradients of the positional parameters are measured separately (Table 3 in the main text).

Table S9: Cosine similarity between the gradients of the two teachers over the parameters of the image encoder, with the weights fixed at each checkpoint (mean over 100 batches; the fraction of batches with a negative cosine in parentheses).

Scheme	5k	10k	15k	20k	25k
None	$+0.13$ (24%)	$+0.02$ (54%)	-0.05 (65%)	$+0.08$ (43%)	-0.05 (66%)
RoPE (NAF)	-0.02 (56%)	-0.08 (65%)	-0.07 (63%)	-0.13 (76%)	-0.15 (81%)
RoPE (learned freq.)	-0.06 (62%)	-0.10 (65%)	-0.04 (60%)	-0.17 (79%)	-0.16 (81%)
Scalar	$+0.03$ (49%)	-0.17 (80%)	-0.24 (88%)	-0.25 (89%)	-0.26 (90%)
Vector	-0.09 (59%)	-0.34 (95%)	-0.28 (87%)	-0.34 (91%)	-0.40 (97%)

Table S10: Reconstruction loss for each teacher at 25k steps. "Single" denotes the model trained with that teacher alone, while "Multi" denotes the model trained with both teachers. The loss is the mean cosine distance ($\times 100$) across 256 ImageNet validation images (one per randomly selected class), which are identical across all models. Lower is better.

	DINOv3		SigLIP2	
Scheme	Single	Multi	Single	Multi
None	13.22	13.59	43.18	43.49
RoPE (NAF)	10.21	10.72	42.23	42.47
RoPE (learned freq.)	10.21	10.77	42.02	42.32
Scalar	10.18	10.61	42.12	42.54
Vector	10.18	10.61	42.06	42.52

Gradients at checkpoints. Table S9 reports the cosine similarity between the gradients of the two teachers for all five schemes, evaluated at checkpoints from 5k to 25k steps with the weights fixed. The gradients are computed over 400 ImageNet training images with a batch size of 4. Under every scheme, the cosine similarity is lower at 25k steps than at 5k steps, and under the four schemes

with a positional encoding, the fraction of batches with a negative cosine similarity rises to 81–97%. At a stationary point of the training objective (the mean of the two losses), the gradients of the two teachers sum to zero. Under every scheme, the multi-teacher model reconstructs each teacher with a higher loss than the model trained on that teacher alone (Table S10), so it settles at a compromise between the two teachers rather than fitting either one as well as a single-teacher model. At this compromise, the two gradients sum to zero without vanishing and therefore point in opposite directions, so a negative cosine similarity late in training is expected under any scheme. At 25k steps, the cosine similarity is most negative for VECTOR and SCALAR, whose Cityscapes gains in Table 1 of the main text are nonetheless positive. A negative cosine similarity late in training is therefore not specific to multiplicative coupling.

Table S11: Cosine similarity between the gradients of the two teachers by parameter group at 25k steps, with the weights fixed (mean over 100 batches; the fraction of batches with a negative cosine in parentheses). The groups are those of Table 3 in the main text.

Scheme	Encoder 1×1	Encoder 3×3	Positional params
None	$+0.22$ (11%)	-0.14 (73%)	–
RoPE (NAF)	-0.12 (68%)	-0.17 (82%)	–
RoPE (learned freq.)	-0.16 (72%)	-0.17 (83%)	-0.58 (91%)
Scalar	-0.47 (100%)	-0.24 (86%)	-0.92 (100%)
Vector	-0.38 (94%)	-0.42 (97%)	-0.91 (100%)

Table S11 breaks down the cosine similarity at 25k steps by parameter group, as Table 3 of the main text does for steps 50–99. Under every scheme with a positional encoding, the cosine similarity is negative in both the 1×1 and 3×3 layers. The gradients of the positional parameters of SCALAR and VECTOR have a cosine similarity of at least $+0.79$ during steps 50–99 but are nearly opposite at 25k steps, with a cosine similarity of -0.91 or lower. Under the four schemes with a positional encoding, the negative cosine similarity at the end of training therefore extends across every parameter group, whereas the early conflict under RoPE is localized in the 3×3 layers of the image encoder. The nearly opposite gradients on the position bias of SCALAR also match §E: the two teachers pull the position bias in opposite directions, and the learned position bias lies between the two single-teacher position biases.

Table S12: Projections applied by PCGrad during the first 100 steps of multi-teacher training. Each step has two candidate projections, one per teacher, so the total is out of 200.

Scheme	Steps 0–49	Steps 50–99	Total
None	0	0	0
RoPE (NAF)	4	38	42
RoPE (learned freq.)	4	44	48
Scalar	0	0	0
Vector	0	0	0

PCGrad projections. Table S12 reports the number of projections that PCGrad applied during the first 100 steps of each scheme's PCGrad run. PCGrad projects only under the two RoPE schemes, mostly in steps 50–99, when the cosine similarity in Figure 3 of the main text falls toward zero and turns negative. In these steps, PCGrad therefore acts only on the early conflict under RoPE. However, PCGrad projects at only 38% and 44% of steps 50–99, fewer than the 76% of steps at which the 3×3 layers conflict (Table 3 of the main text). This is because PCGrad checks the sign of the inner product computed over all parameters, including the 1×1 layers, where the gradients of the two teachers remain aligned on average.