Zero-Ablation Overstates Register Content Dependence in DINO Vision Transformers

Supplementary Material

Felipe Parodi[1] Jordan K. Matelsky[1,2] Melanie Segado[1]

[1]University of Pennsylvania [2]Johns Hopkins Applied Physics Laboratory

{fparodi, matelsky, msegado}@upenn.edu

Table 1. ViT-B task × ablation matrix (layer 11). CLS: probe top-1 (%). Corr: correspondence (%). Seg: segmentation mIoU (%). SPair: PCK@0.1 (%). Δ: change from Full.

		CLS	Corr	Seg	SPair
DINOv2-B	Full	78.7	72.9	72.0	41.2
	Zero CLS	0.1 -78.6	58.9 -14.0	46.1 -25.9	21.3 -19.9
DINOv2-B +reg	Full	74.5	71.2	72.3	41.1
	Zero CLS	0.1 -74.4	70.4 -0.8	72.3 0.0	41.2 +0.1
	Zero Reg	55.2 -19.3	63.3 -7.9	64.1 -8.2	28.8 -12.3
DINOv3-B	Full	73.3	77.1	83.4	37.9
	Zero CLS	0.1 -73.2	79.5 +2.4	82.8 -0.6	37.8 -0.1
	Zero Reg	36.8 -36.5	61.3 -15.8	59.6 -23.8	19.1 -18.8

Table 2. Segmentation probe mIoU (%, ViT-S, VOC2012 val, layer 11, mean ± std over 3 seeds).

Model	Ablation	mIoU
DINOv2	Full	70.8 ± 0.10
	Zero CLS	33.7 ± 0.07
DINOv2+reg	Full	71.3 ± 0.07
	Zero CLS	70.7 ± 0.11
	Zero Reg	61.7 ± 0.14
DINOv3	Full	78.5 ± 0.00
	Zero CLS	78.5 ± 0.04
	Zero Reg	47.6 ± 0.08

1. Extended Results

The ViT-B task × ablation matrix (Tab. 1) complements the ViT-S matrix in the main text (**??**). Per-task breakdowns with confidence intervals follow.

Segmentation.

Synthetic correspondence. Table 3 provides per-condition correspondence with bootstrapped CIs. Register zeroing reduces correspondence from 69–79% to 58–64% (ViT-S; Tab. 3), visible as collapsed match patterns in **??**.

Tolerance sensitivity. Ablation patterns are stable across tolerance thresholds (Tab. 4). DINOv3 shows the largest register-zeroing drop at all thresholds (-18.2 to -21.1 pp).

SPair-71k quantization ceiling. DINOv3 uses 16-pixel patches (14×14 grid, 196 tokens) vs. DINOv2's 14-pixel patches (16×16 grid, 256 tokens). An oracle quantization test yields ceilings of 97.9% (16×16 grid) and 96.5% (14×14 grid)—a 1.4 pp gap that accounts for only a fraction of the 6.6 pp absolute PCK difference, so grid resolution is not the dominant factor. ViT-B SPair results follow the same asymmetry (Tab. 1).

SPair-71k correspondence. The SPair-71k results table is in the main text (**??**).

2. Controls and Statistical Tests

Random-patch negative control. Zeroing 4 random patch tokens (5 seeds) causes ≤ 1 pp CLS drop for ViT-S and ≤ 2.3 pp for ViT-B (Tab. 5), vs. -18.9 / -36.6 pp for register zeroing.

Mean-substitution control. Replacing registers with per-layer dataset-mean activations (5,000 images; **??** in main text) has negligible effect (-0.3 pp and $+0.1$ pp) vs. -18.9

Table 3. Patch correspondence (ViT-S, layer 11, 2,000 pairs). GT acc: ground-truth accuracy (%). Cycle: cycle consistency (A→B→A). Bootstrapped 95% CI in brackets.

Model	Ablation	GT acc	Cycle
DINOv2	Full	72.0 [71.4, 72.7]	0.468
	Zero CLS	56.1 [55.2, 57.1]	0.411
DINOv2+reg	Full	69.1 [68.4, 69.9]	0.454
	Zero CLS	68.3 [67.6, 69.2]	0.450
	Zero Reg	64.3 [63.4, 65.2]	0.433
DINOv3	Full	78.9 [78.2, 79.8]	0.477
	Zero CLS	78.2 [77.4, 79.1]	0.477
	Zero Reg	57.8 [56.5, 59.2]	0.331

Table 4. Correspondence accuracy (%) at different tolerance thresholds (ViT-S, layer 11). Tol = 0: exact match; 1: ±1 patch (default); 2: ±2 patches.

Model	Ablation	Tol = 0	Tol = 1	Tol = 2
DINOv2	Full	38.5	72.0	80.3
	Zero CLS	28.0	56.1	65.2
DINOv2 +reg	Full	37.5	69.1	77.1
	Zero CLS	36.8	68.3	76.4
	Zero Reg	34.0	64.3	72.7
DINOv3	Full	45.2	78.9	85.3
	Zero CLS	45.0	78.2	84.0
	Zero Reg	27.0	57.8	66.4

and −36.6 pp under zeroing. Mean-substituting CLS yields 0.1% (same as zeroing), confirming CLS carries all image-specific class signal. Classification accuracy under mean-substitution is insensitive to calibration set size: varying N from 100 to 5,000 images changes accuracy by ≤0.1 pp for both DINOv2+reg (67.0–67.1%) and DINOv3 (62.0–62.1%), indicating the control does not depend on precise mean estimates.

Noise-substitution and register-shuffling controls. *Noise-substitution* replaces register outputs at each layer with Gaussian noise matched in per-dimension mean and variance (calibrated on 5,000 images), preserving marginal statistics but destroying register-specific structure. *Register shuffling* permutes register activations across images within each batch (independently at each layer), preserving real activation structure but breaking image-specific routing. All three controls preserve performance across classification, correspondence, and segmentation (see **??** in main text for complete results).

Statistical significance. Per-image or per-token outcome differences are tested with sign-flip permutation tests (10,000 permutations; Tab. 6). The two dissociation comparisons are significant at $p < 0.001$; the scale comparison ($p = 0.80$) confirms register dependence is consistent across ViT-S and ViT-B.

Table 5. Random-patch negative control: CLS top-1 (%) when zeroing 4 random patch tokens (ViT-S and ViT-B, layer 11, 5 seeds). Δ: change from Full baseline (**??** and Tab. 1). Reg-zero Δ reproduced for comparison.

Model	Condition	CLS (%)	Δ	Reg-zero Δ
DINOv2	Rand. patch	72.4 ± 0.10	−0.8	—
DINOv2+reg	Rand. patch	66.3 ± 0.03	−1.0	−18.9
DINOv3	Rand. patch	61.5 ± 0.05	−0.5	−36.6
DINOv2-B	Rand. patch	78.4 ± 0.07	−0.3	—
DINOv2-B+reg	Rand. patch	74.5 ± 0.04	0.0	−19.3
DINOv3-B	Rand. patch	71.0 ± 0.06	−2.3	−36.5

Table 6. Paired permutation test p-values for headline comparisons (10,000 permutations, two-sided).

Comparison	Observed Δ	p
DINOv3 vs. DINOv2+reg Δ_{ZeroReg} (CLS)	17.7 pp	< 0.001
CLS-zeroing buffering (Seg, with vs. without regs)	10.7 pp	< 0.001
ViT-S vs. ViT-B register dependence (DINOv3 Δ_{ZeroReg})	0.1 pp	0.80 (ns)

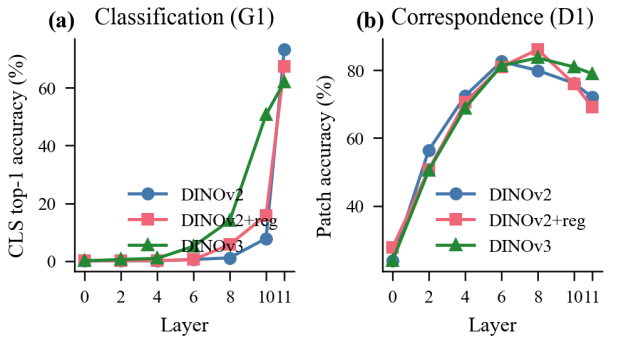

Figure 1. **Task performance across transformer layers. (a)** CLS classification (linear probe, 50 epochs, 1 seed). Classification emerges at layers 10–11. **(b)** Patch correspondence (tolerance = 1). Correspondence peaks at layers 6–8 then declines, except DINOv3 which maintains 78.9% at layer 11. ViT-S models.

3. Representation Geometry

Effective rank across layers reveals when patch compression and register dependence emerge (Tabs. 7 and 8 and Fig. 1; see also **??**). DINOv3 is already compressed at layer 6 (effective rank 6.4 vs. 32.3 for DINOv2), yet register dependence for classification emerges only at layers 10–11. In DINOv3, register zeroing *improves* CLS accuracy at intermediate layers before becoming strongly detrimental at the final layers, suggesting registers become structurally expected only as classification information consolidates.

4. Mechanistic Analysis

Attention flow across layers. **??** traces how attention mass distributes between token types at each of the 12 transformer layers. In DINOv2 (no registers), CLS self-attention dominates early layers then declines. In both register models,

Table 7. Effective rank (median ± std) at layers 6 and 11 (ViT-S, 500 ImageNet images). Layer 11 values match the Full rows of Tab. 8.

Model	Layer 6	Layer 11
DINOv2	32.3 ± 6.3	13.5 ± 4.5
DINOv2+reg	41.7 ± 7.4	8.7 ± 3.3
DINOv3	6.4 ± 7.0	4.0 ± 1.2

Table 8. Median Gram statistics (ViT-S, layer 11, 2,000 images). DINOv3 entropy omitted: its 14×14 patch grid yields a 196×196 Gram matrix vs. 256×256 for DINOv2 models, making eigenspectrum entropy incomparable.

Model	Ablation	Erank	Entropy
DINOv2	Full	13.5	2.61
	Zero CLS	18.1	2.90
DINOv2+reg	Full	8.7	2.16
	Zero CLS	9.8	2.29
	Zero Reg	11.3	2.43
DINOv3	Full	4.0	—
	Zero CLS	4.5	—
	Zero Reg	5.1	—

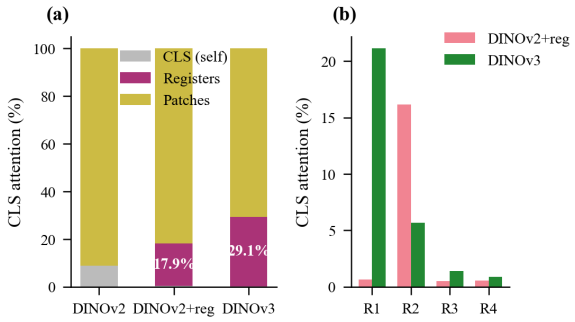

Figure 2. **CLS attention (ViT-S, last layer, 200 images). (a)** CLS attention fraction per token type. DINOv2+reg: 17.9% to registers; DINOv3: 29.1%. **(b)** Per-register breakdown. This routing structure is maintained under all plausible register replacements.

register attention share builds *gradually* from mid-layers: DINOv2+reg stabilizes at ∼20% CLS→registers by layer 6, while DINOv3 ramps more steeply to 28.7% at layer 11. This gradual ramp contrasts with the abrupt emergence of classification accuracy at layers 10–11, dissociating attention routing from functional dependence.

Attention pattern analysis. **??** in the main text compares attention patterns under register zeroing vs. mean-substitution across all 12 layers. At ViT-S scale, register zeroing yields last-layer JS divergence of 0.144 (DINOv2+reg) and 0.177 (DINOv3), while mean-substitution yields 0.005 and 0.001 respectively—a 29× and 253× gap. Divergence is identically zero at layer 0 (same input) and amplifies across layers, confirming that the distributional shift from zero vectors cascades rather than remaining a single-layer effect. CLS zeroing produces much smaller divergence (0.026 and 0.010), consistent with CLS being a downstream reader rather than a structural element whose removal causes large distributional shift.

Per-register dose-response. The register whose removal causes the greatest attention disruption matches the register with the highest decodable class information: DINOv2+reg R2 (JS = 0.062, vs. 0.020–0.025 for other registers) and DINOv3 R3 (JS = 0.076, vs. 0.006–0.022). Note that zeroing individual registers is also a distribution-shifting intervention, so these results should be interpreted as measuring sensitivity to distributional shift rather than functional dependence on register content.

Scale consistency. At ViT-B scale, the pattern replicates: register zeroing yields last-layer JS of 0.232 (DINOv2-B+reg) and 0.183 (DINOv3-B), while mean-substitution yields 0.009 and 0.002 (**??**a, lighter lines).

The attention-flow and PCA-projection figures are in the main text (**????**).

Register token analysis. Initial per-register lesions and probes suggested register specialization. However, the substitution controls in **??** show that these patterns need not reflect functional dependence: class information is decodable from individual registers, yet models do not require it for any measured downstream task. We report these analyses as exploratory and descriptive; they characterize representational structure but not functional necessity.

5. ViT-B Scale Validation

We replicate the zero-ablation experiments at ViT-B scale: DINOv2-B (86.6M params), DINOv2-B+reg (with four register tokens), and DINOv3-B (85.7M params). The full ViT-B ablation matrix is in Tab. 1; replacement controls are in the main text (**??**).

Task-level replication. ViT-B absolute accuracies are higher but ablation deltas are nearly identical to ViT-S. Classification: DINOv3-B loses −36.5 pp [-38.2, -34.6] under register zeroing (vs. −36.6 at ViT-S); DINOv2-B+reg loses −19.3 pp [-21.2, -17.5] (vs. −18.9). Correspondence: CLS zeroing degrades DINOv2-B by −14.0 pp but is negligible with registers; register zeroing hits DINOv3-B hardest (−15.8 vs. −7.9 pp). SPair-71k: same pattern (−18.8 pp DINOv3-B vs. −12.3 pp DINOv2-B+reg). Segmentation: CLS zeroing drops DINOv2-B by −25.9 pp but ≤ 0.6 pp with registers; register zeroing again hits DINOv3-B hardest (−23.8 vs. −8.2 pp). The Gram compression pattern replicates: effective rank 20.1 → 13.9 → 6.6 (3.0× compression from DINOv2-B to DINOv3-B, vs. 3.4× at ViT-S).

Attention and per-register patterns. ViT-B attention routing mirrors ViT-S: DINOv2-B+reg routes 25.6% of CLS attention to registers (R4 dominant at 23.5%); DINOv3-B routes 30.6% (R1 largest at 15.4%, more distributed). Per-register zero-lesions show a distributed pattern in DINOv3-B (R3: −1.7 pp, R2: −1.5 pp, R1: −1.1 pp, R4: −0.9 pp) while DINOv2-B+reg shows minimal individual effects (≤ 1.1 pp); however, zeroing individual registers is also a distribution-shifting intervention (Sec. 4). The identity of the dominant attention register shifts from R2 (ViT-S) to R4

(ViT-B), suggesting that register differentiation is not fixed across scales but emerges from training dynamics.

6. Experimental Details

Feature extraction. All features are extracted using HuggingFace `transformers` (`facebook/dinov2-small` and `facebook/dinov2-with-registers-small` for ViT-S; `facebook/dinov2-base` and `facebook/dinov2-with-registers-base` for ViT-B). DINOv3 models are loaded via `torch.hub` with locally cached weights (`dinov3_vits16` and `dinov3_vitb16`). Input images are resized to 224×224 and normalized with ImageNet statistics. We extract features from the final block output after the model's terminal LayerNorm (layer 11 of 12).

Ablation hooks. PyTorch forward hooks replace the relevant token positions after every block output, starting from block 1. Zero-ablation substitutes $\mathbf{0} \in \mathbb{R}^d$; mean-substitution uses per-layer dataset-mean activations calibrated on 5,000 images; noise-substitution uses per-layer Gaussian noise matched in mean and variance; register shuffling permutes activations across images within each batch independently at each layer.

Linear probe training. Single linear layer ($d \to K$). SGD with momentum 0.9, weight decay 0.1, learning rate 0.01 with cosine annealing, 100 epochs, batch size 256. Stratified 80/20 split (40,000 / 10,000 images, seed 42).

Segmentation probe training. Single linear layer ($d \to 21$) per patch token; masks downsampled to the patch grid via nearest-neighbor interpolation. AdamW, weight decay 10^{-2}, learning rate 10^{-3}, constant, 100 epochs. Per-pixel cross-entropy, ignoring void (index 255).

kNN retrieval. 2,000 ImageNet val images, each producing two augmented views (`RandomResizedCrop`, `ColorJitter`, `RandomHorizontalFlip`). Cosine similarity; R@1 with 1,000 bootstrap resamples.

Patch correspondence. Same augmentation pipeline as kNN, with crop coordinates recorded for ground-truth spatial correspondence. A match is correct if the nearest-neighbor patch falls within 1 patch of ground truth; cycle consistency requires the chain A→B→A to return to the exact original patch.

SPair-71k evaluation. Images resized to 224×224; source keypoints mapped to the patch grid; correspondence predicted via cosine similarity. PCK@0.1: correct if Euclidean distance $< 0.1 \times \max(h_{\text{bbox}}, w_{\text{bbox}})$. Only mutually visible keypoints evaluated; 1,000 bootstrap resamples.

Attention analysis. Attention weights from the final block (`attn_implementation="eager"` for HF models). CLS row averaged over heads, then over 200 images. Layer-sweep probes use 50 epochs / 1 seed (lower absolute accuracy than main-text probes; relative layer-wise patterns are stable).

Attention rewiring analysis. Jensen–Shannon divergence is computed between full and ablated attention distributions at every layer and head (1,000 images, all 12 layers). Per-head JS values are averaged across heads and images to produce per-layer divergence curves.

Compute. Single NVIDIA RTX 4090 (24 GB); total GPU time ≈12–15 hours (feature extraction across 6 models $\times$ multiple ablation conditions $\times$ 4 datasets, probe training, attention analysis, and mean-activation calibration).

References