WHAT ARTIFACT TOKENS IN VISION TRANSFORMERS ARE, AND WHAT THEY DO

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers develop a handful of high-norm "artifact" tokens that occupy low-information image regions and absorb a disproportionate share of attention. Register tokens were introduced to eliminate them, and a subsequent audit reported that the phenomenon does not generalise uniformly across architectures. We present a measurement study across 15 pretrained checkpoints and 10 datasets spanning 9 visual domains. First, artifact formation is determined by model weights, not input statistics, as varying the fraction of low-information patches over a $9.1\times$ range changes artifact strength with a coefficient of variation of only 0.030–0.102, while leaving the peak-artifact layer exactly invariant for every model on every dataset, including histopathology, chest radiography, and satellite imagery. Second, the behaviour emerges with scale, appearing between 22M and 87M parameters on matched DINOv2 checkpoints. Third, register tokens work by capturing the read target. Adding four registers collapses attention mass on high-norm patch tokens from 0.382 to 0.007, renders those patches causally inert under ablation, and makes the registers load-bearing instead. Fourth, for image–text contrastive models, removing eight of roughly 200 tokens costs 15–21 accuracy points under every pooling operator tested, with matched random-ablation controls showing no effect. Fifth, the sign of the ablation effect depends on the training objective rather than the evaluation protocol. It remains invariant across a band of layers, three evaluators, and neighbourhood sizes spanning two orders of magnitude, while exceeding matched random controls by 5–$20\times$. This resolves an apparent protocol dependence in which pooled and per-token evaluation yield opposite-signed conclusions about the same intervention.

1 INTRODUCTION

A Vision Transformer (Dosovitskiy et al., 2021) partitions an image into patches and represents each as a token vector refined across layers by attention (Vaswani et al., 2017). Darcet et al. (2024) observed that in most large ViTs a handful of these tokens acquire norms one to two orders of magnitude larger than the rest, and that these artifact tokens occupy the least informative image regions. Adding a small number of dedicated register tokens removes the visible artifacts and improves dense prediction and unsupervised object discovery. Three follow-ups have sharpened the finding. Jiang et al. (2026) identified a sparse set of neurons responsible for concentrating high-norm activations and used them to construct a training-free mitigation. Fesser et al. (2026) unified register tokens with gated attention (Qiu et al., 2025). Baxevanakis et al. (2026) audited the phenomenon across architectures and found that several original claims do not extend universally, an inconsistency that remains unexplained. We step back from mitigation and ask what is actually happening. The paper is organised around two questions that are addressed through five findings (F1–F5). Section 4–Section 6 ask what artifact tokens are: what determines that they form, when the behaviour appears, and what register tokens actually change. Section 7–Section 8 ask what they do: whether they carry information, and what determines the sign of the effect when they are removed. Our contributions are measurements rather than architecture, and two of them constrain interpretations that would otherwise be natural to adopt. Figure 1 shows why this matters. Ablating the same eight tokens from DINOv2 raises pooled k-NN accuracy by 8.6 points but lowers per-token segmentation mIoU by 2.5. Both are standard protocols with matched controls, so either result alone would be internally consistent yet misleading. We resolve this apparent contradiction in Section 8.

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.

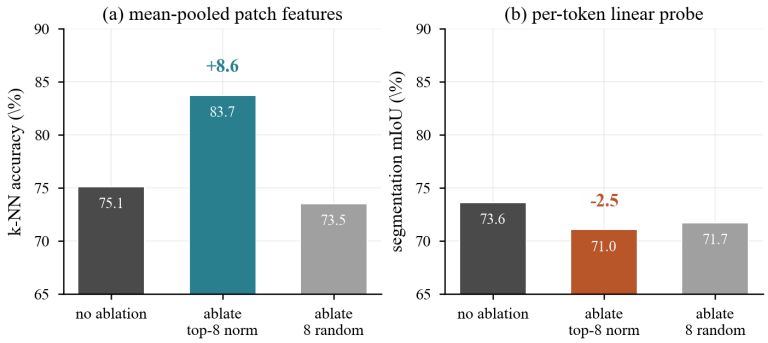


Figure 1: Ablating the eight highest-norm patch tokens (3% of 256) from DINOv2 L/14 at its peak-artifact layer. **(a)** With mean-pooled features the intervention appears to improve the representation by 8.6 points ($\approx 7\sigma$), while a matched random control shows no gain. **(b)** A per-token linear probe on the same frozen features reverses the sign: mIoU falls by 2.53 points, and high-norm ablation becomes indistinguishable from random.

For what they are, our first three findings concern what determines whether artifact tokens exist at all. Their formation is fixed by the model weights rather than by the input distribution, as shown by results across 10 datasets spanning 9 visual domains and a $9.1\times$ range in low-information content. Artifact strength varies with a coefficient of variation of 0.030–0.102, while the layer at which artifacts peak does not move at all (F1, Section 4). What the weights encode is, in turn, a function of scale, with attention mass received by artifact tokens rising from 0.008 at 22M parameters to 0.382 at 87M. The behaviour therefore emerges within a narrow band rather than growing smoothly (F2, Section 5). Given that artifacts are weight-determined, the natural question is what register tokens change. Using two independent methods on three matched scale pairs, we find that they do not remove the underlying pressure but redirect it, capturing attention that would otherwise fall on patch tokens and becoming load-bearing in their place (F3, Section 6). For what they do, having established what these tokens are, we ask what they carry. The answer depends on the training objective rather than on artifact severity. For image–text contrastive models, removing the 4% of tokens with the highest norms costs 15–21 accuracy points under every pooling operator tested, and the geometry of the resulting representation shows images collapsing toward one another. Self-distillation models are unaffected by the identical intervention, even when their artifacts are three times stronger (F4, Section 7). Finally, we resolve the protocol dependence of Figure 1 by showing that the sign of the ablation effect is set by the training objective, not by the evaluation protocol, and holds across a contiguous band of layers, three evaluators, and neighbourhood sizes from 1 to 100 (F5, Section 8). Section J reports a further negative result obtained without training: extending gated attention from per-query to per-query–key gating offers no practical capacity advantage over an additive query-plus-key gate.

2 Background and Related Work

Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be the input tokens to an attention layer, projected into queries, keys and values $q_i = W_Q x_i$, $k_i = W_K x_i$, $v_i = W_V x_i$. Attention produces an output $O_i \in \mathbb{R}^{d_v}$ at each position,

$$O_i = \sum_j a_{ij} v_j, \qquad a_{ij} = \mathrm{softmax}_j\left(\frac{q_i \cdot k_j}{\sqrt{d_k}}\right), \qquad \sum_j a_{ij} = 1, \tag{1}$$

projected back to $\mathbb{R}^d$ by W_O and added to the residual stream; in the multi-head case this is computed in H subspaces, and we suppress the head index except in Section J. The constraint $\sum_j a_{ij} = 1$ is the origin of the phenomenon we study. Since O_i is a convex combination of value vectors, a head with nothing relevant to read cannot abstain. It must still emit a weighted average into the residual stream. The available workaround is to designate some token as a destination for surplus attention. In language models this is typically a fixed position, producing an attention sink (Xiao et al., 2024); proposed remedies relax the constraint in various ways (Qiu et al., 2025; Ye et al., 2025). Vision Transformers have no canonical sink position, so the destination must be discovered from the image

itself, which is why our read score (Section 3) is defined over a per-image outlier set rather than a fixed index. Darcet et al. (2024) characterised these tokens in ViTs and removed them with dedicated register tokens, at the cost of retraining from scratch. Jiang et al. (2026) located the responsible neurons and built a training-free mitigation without addressing why the tokens form; Fesser et al. (2026) showed gating and registers are complementary, leaving the mechanism partly open; and Baxevanakis et al. (2026) found the original claims do not extend universally across architectures, reporting the variation without explaining it. Two gaps motivate this work. First, existing studies do not establish what determines whether artifacts form. Second, they do not distinguish tokens that are written to from tokens that are read from, a distinction necessary for measuring what registers actually change (Section 6).

3 METHOD

All measurements are inference-only on frozen public checkpoints from timm (Wightman, 2019). No training was performed at any point. Token norms grow substantially with depth, so a global threshold is invalid. Let $h_i^{(\ell)} \in \mathbb{R}^d$ denote the residual-stream representation of patch token i at the output of block ℓ, for $i = 1, \ldots, P$ patch tokens (prefix tokens such as CLS and registers are excluded throughout). We identify outliers by a robust z-score,

$$z_i^{(\ell)} = \frac{\|h_i^{(\ell)}\|_2 - \text{median}_j \|h_j^{(\ell)}\|_2}{1.4826 \cdot \text{MAD}_j \|h_j^{(\ell)}\|_2}, \qquad \text{outlier} \iff z_i^{(\ell)} > 5, \tag{2}$$

with median and MAD taken over the P patch tokens of a single image at a single layer. The constant $1.4826 = 1/\Phi^{-1}(0.75)$ makes $1.4826 \cdot \text{MAD}$ a consistent estimator of the standard deviation under normality, so the threshold retains its conventional interpretation; we use the MAD rather than $\hat{\sigma}$ because the latter is inflated by the very outliers we wish to detect.

Artifact strength: For an image with patch-token representations $h^{(\ell)}$, artifact strength at layer ℓ is the ratio of the largest to the median token norm, averaged over B (the number of images per batch) images:

$$\rho_{\text{art}}^{(\ell)} = \frac{1}{B} \sum_{b=1}^{B} \frac{\max_i \|h_i^{(\ell,b)}\|_2}{\text{median}_i \|h_i^{(\ell,b)}\|_2}, \qquad \rho_{\text{art}} = \max_\ell \rho_{\text{art}}^{(\ell)}, \tag{3}$$

and $\ell^\star = \arg\max_\ell \rho_{\text{art}}^{(\ell)}$ is the peak-artifact layer. Equation (2) is the standard robust z-score (Iglewicz & Hoaglin, 1993); our contribution is to compute it per image and per layer, which is necessary because norms grow by an order of magnitude with depth. Equation (3) differs from prior practice. Darcet et al. (2024) use a fixed absolute cutoff and Baxevanakis et al. (2026) a 98th-percentile cutoff over a reference subset, neither of which is scale-free. The max-over-median ratio removes that dependence, which is what permits the cross-dataset comparison in Table 1. Threshold sensitivity is verified in Section B: varying the outlier threshold over $z \in [3, 20]$ changes read score by less than 0.02 and leaves the peak layer unchanged. Throughout, P is the number of patch tokens (196 at patch size 16, 256 at patch size 14), and L the number of blocks. We zero the k highest-norm patch tokens at $\ell^\star$ via a forward hook and measure the change in the downstream representation, using $k = 8$ throughout. The random-ablation control, zeroing k uniformly sampled patch tokens, is essential: without it the effect of removing k tokens cannot be separated from the effect of removing those k. We average it over three seeds and report it alongside every ablation result. For models without artifacts $\ell^\star$ is undefined, since $\arg\max$ over a flat profile returns an arbitrary index; we report it only when $\rho_{\text{art}} > 3.0$ and otherwise use mid-depth. This is fixed independently of the $z > 5$ criterion so that the random-ablation control is exactly matched in size; Section D verifies that the sign and approximate magnitude of every result are stable for k from 1 to 16, and that ablating exactly the per-image outlier set instead gives the same answer.

Measurement layer: Artifact strength is compressed by the final LayerNorm: measured at $\ell^\star$ versus at the output, DINOv2 B/14 gives 57.1 versus 4.0, CLIP B/16 22.9 versus 5.8, SigLIP B/16 19.7 versus 5.0. Reported artifact strength must state the measurement layer; we use $\ell^\star$ throughout. Note that the layer of maximal read score sits two to three blocks deeper than $\ell^\star$ in every strong-artifact model (Section B), so the two quantities peak at different depths.

Read score: The read score s_{read} is the fraction of total attention mass received by outlier tokens at a given layer, averaged over queries, heads and images. Measuring mass received by a token is

Table 1: Artifact strength ρ_{art}, ordered by low-information content (top row). Peak layer in parentheses; CV is the coefficient of variation across datasets. The rightmost column gives the set of distinct peak layers observed, a singleton for every model. $n = 800$ images per dataset.

	DTD	Flow.	Food	Pets	Path	Chest	Blood	Airc.	Derm.	Euro			
% flat	6.4	9.3	10.5	12.3	16.1	19.5	33.2	41.0	45.0	58.6	mean	CV	layers
CLIP B/16	24.0 (7)	23.4 (7)	24.2 (7)	22.7 (7)	23.9 (7)	24.4 (7)	25.1 (7)	21.2 (7)	23.4 (7)	25.3 (7)	23.8	0.047	{7}
DINOv2 B/14	56.0 (8)	57.8 (8)	60.2 (8)	57.0 (8)	55.9 (8)	57.9 (8)	55.4 (8)	56.3 (8)	59.1 (8)	60.6 (8)	57.6	0.030	{8}
SigLIP B/16	21.7 (6)	20.2 (6)	20.7 (6)	19.6 (6)	22.9 (6)	21.8 (6)	23.6 (6)	21.2 (6)	23.0 (6)	24.1 (6)	21.9	0.065	{6}
DeiT-III B/16	2.3 (10)	2.7 (10)	2.9 (10)	2.3 (10)	2.2 (10)	2.8 (10)	2.4 (10)	2.2 (10)	2.4 (10)	2.2 (10)	2.4	0.102	{10}

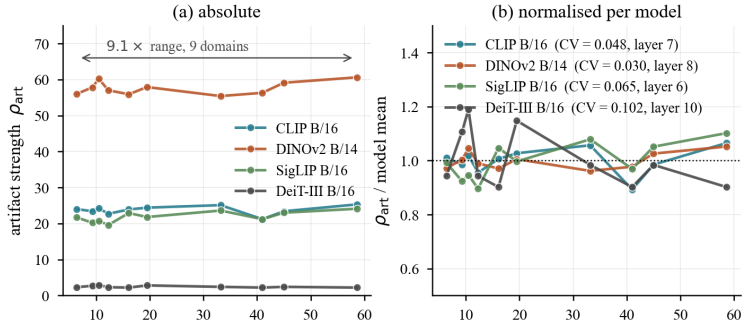


Figure 2: Artifact strength: **(a)** Absolute values; the three strong-artifact models occupy well-separated bands. **(b)** Each model divided by its own mean, isolating the input effect.

standard in the attention-sink literature, where it is computed for a fixed position (Xiao et al., 2024; Qiu et al., 2025). Vision transformers have no canonical sink position and the absorbing tokens differ between images, so we compute mass received by the outlier set of Equation (2), rediscovered per image and layer. Separating which tokens have large norms from which tokens receive attention is what makes Section 6 possible. In a register-trained model, no high-norm patch tokens remain to measure, so the transfer of function is visible only in the attention statistics.

Checkpoints and validation: We evaluate 15 checkpoints spanning contrastive, self-distillation, supervised and reconstruction objectives; the full sweep is Table 11 in Section F. Two observations shape what follows. Artifact strength and read score are tightly coupled: no model exhibits strong artifacts without a high read score. And the checkpoints separate into two clusters, high-artifact/high-read and low-artifact/low-read, with no intermediate cases, which is why we treat artifact severity as a single axis. Before drawing conclusions we verified that the suite reproduces three published findings: DINO is the artifact-free exception (ρ_{art} 1.4–1.6 versus CLIP 23.2, DINOv2 57.9); outliers occupy low-information patches (they sit at the 19th percentile of within-image local variance); and artifacts emerge above a scale threshold (Section 5).

4 F1: Artifact formation is weight-driven

We hypothesize that if artifacts form because the model requires low-information patches to write into, then image distributions containing more such content should produce stronger artifacts. This is the natural reading of Darcet et al. (2024)'s placement observation, and would offer a candidate explanation for the cross-architecture variation reported by Baxevanakis et al. (2026). We assemble 10 datasets spanning 9 visual domains and quantify low-information content as the percentage of 16×16 patches whose mean per-channel standard deviation falls below 0.02 (top row of Table 1). The panel spans from 6.4% (textures) to 58.6% (satellite imagery), a $9.1\times$ ratio, and deliberately includes domains far outside the ImageNet-like distribution on which these models were pretrained. Table 1 reports artifact strength and peak layer for four models on all ten datasets; Figure 2 plots the same data, with the right panel dividing each model by its own mean to remove between-model scale differences and isolate the input effect. The resulting lines are flat near 1.0 across the entire range.

Table 2: Read score s_{read} on matched DINOv2 checkpoints. Read behaviour emerges between 22M and 87M parameters and is suppressed by registers at every scale.

Model	Params	s_{read} (no registers)	s_{read} (+4 registers)
ViT-S/14	22M	0.008	0.010
ViT-B/14	87M	**0.382**	0.007
ViT-L/14	304M	**0.430**	0.011

Quantitatively, artifact strength varies with coefficient of variation 0.030–0.102 across a $9.1\times$ range in the independent variable. More strikingly, the peak-artifact layer is identical for every model on every dataset: layer 7 for CLIP, 8 for DINOv2, 6 for SigLIP, 10 for DeiT-III, across every domain in the panel. Not only does the magnitude of the phenomenon fail to track the input, its location in the network does not move at all. Figure 3 shows the same conclusion qualitatively. Pearson correlations between flat-patch fraction and artifact strength are 0.096 for CLIP ($p = 0.79$), 0.338 for DINOv2 ($p = 0.34$), 0.694 for SigLIP ($p = 0.026$), and -0.527 for DeiT-III ($p = 0.12$). SigLIP's correlation is nominally significant but does not survive Bonferroni correction across four models ($\alpha = 0.0125$). Its effect size is small, ranging from 19.6 to 24.1, a 23% spread across a $9.1\times$ range in the input. The signs are also inconsistent, with DeiT-III showing a negative correlation. These inconsistent signs at small magnitudes are more consistent with noise than with a systematic mechanism. A plateau across datasets could in principle reflect saturation rather than input independence, since every dataset in the panel supplies more low-information patches than a model needs. We rule this out with synthetic inputs containing no flat patches at all: artifacts still form at 80–95% of panel-mean strength and at the same peak layer (Section C). Artifact formation is a property of the trained weights. Low-information regions determine where artifacts land, confirming Darcet et al. (2024)'s placement observation, which we independently reproduce, but they do not determine whether or how strongly artifacts form. This finding refines rather than contradicts the existing account and has a practical consequence. Because the input distribution does not modulate the phenomenon, artifact behaviour can be characterised reliably from a single dataset, while cross-dataset variation in reported artifact strength is more plausibly attributed to differences in measurement protocol than to the data themselves. Section H provides a complementary view. The correlation between token norm and local patch variance changes sign with depth at a model-specific layer, offering a scalar summary of when a model begins repurposing patches.

5 F2: Emergence with scale

Having established that artifact behaviour is largely determined by the trained weights, we next ask what aspect of training determines its emergence. Table 2 reports read score on matched DINOv2 checkpoints (Oquab et al., 2024) at three scales, with and without register tokens. The comparison is unusually clean because the architecture family, pretraining data, and objective are held fixed, with only four additional tokens distinguishing the register-trained variants. Read behaviour emerges between 22M and 87M parameters and then largely saturates, rising only from 0.382 to 0.430 across a further $3.5\times$ increase in parameters. The register-trained variants remain near 0.01 at all three scales, including 22M, where the behaviour has not yet emerged. This parallel structure provides a matched control for the scale effect.

6 F3: Registers capture the read target

The register variants in Table 2 show near-zero read scores, but that alone does not establish whether the mechanism was removed or merely relocated. We address this with two measurements that could in principle disagree: one correlational, asking where the model directs attention, and one causal, asking what breaks when specific tokens are removed. Attention mass becomes decisive only when we ask where the mass goes rather than merely noting its absence from patch tokens; ablation then establishes that the receiving tokens are load-bearing. In DINOv2 B/14, attention mass on high-norm patch tokens is 0.382. Adding four registers collapses this mass to 0.007. A full accounting of where it goes (Section I) shows that the destination is not the ordinary patch tokens, whose share is essentially unchanged (0.663 to 0.641), but the registers, which receive 0.304 while CLS

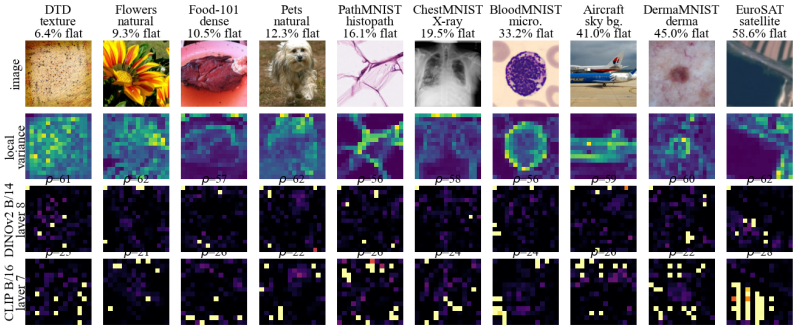


Figure 3: One representative image per dataset, ordered from left to right by increasing low-information content. Rows show the input image, local patch variance, and robust z-scores of token norms at the peak-artifact layer for DINOv2 B/14 and CLIP B/16.

Table 3: Ablation damage as cosine similarity to the unablated representation; lower indicates more damage. Adding registers makes patch tokens inert and registers load-bearing (Plotted in Figure 16).

Model	Top-4 patches	4 registers	4 random patches	CLS
DINOv2 S/14	0.992	—	0.998	0.517
DINOv2 S/14 +reg	0.984	0.934	0.995	0.457
DINOv2 B/14	**0.798**	—	0.955	0.364
DINOv2 B/14 +reg	0.984	**0.880**	0.996	0.768
DINOv2 L/14	**0.823**	—	0.993	0.232
DINOv2 L/14 +reg	0.995	**0.575**	0.996	−0.307

sheds 0.267. Registers therefore relieve the CLS token of its sink role rather than merely displacing attention within the patch grid. Table 3 reports cosine similarity between the ablated and unablated global representation, where lower values indicate more damage. In the base models, top-4 patch ablation is clearly more damaging than random ablation at ViT-B (0.798 vs. 0.955) and ViT-L (0.823 vs. 0.993), while at ViT-S the two are indistinguishable (0.992 vs. 0.998), consistent with the scale threshold of Section 5. In the register variants, patch ablation becomes statistically indistinguishable from random ablation (0.984 vs. 0.996; 0.995 vs. 0.996), while the damage instead appears in the register column. At ViT-L, the registers reach 0.575, exceeding the damage that patch outliers ever caused in the base model. The function therefore did not merely move but consolidated. Registers therefore do not remove the pressure that produces artifacts; they provide a better destination for it. The CLS column adds a complementary observation. Zeroing CLS in DINOv2 L/14 +reg drives the representation to anti-correlation (−0.307), indicating that CLS and registers perform distinct roles. Section H gives a layer-resolved view of the phenomenon in CLIP.

7 F4: Contrastive models depend on artifact tokens

Having established what artifact tokens are and how registers affect them, we ask what they carry. Following standard practice for frozen-feature evaluation (Caron et al., 2021), we report k-NN classification accuracy with $k = 20$ unless otherwise stated. Table 4 reports k-NN classification accuracy on Oxford-IIIT Pet under ablation of the eight highest-norm patch tokens (3–4% of the sequence), with bootstrap 95% confidence intervals. Confidence intervals are non-parametric bootstrap percentile intervals (Efron & Tibshirani, 1993) over 300 resamples of the query set. The split is sharp. The two image–text contrastive models lose 13–14 accuracy points with non-overlapping confidence intervals, while every self-distillation and supervised model shows differences well inside noise. Note that the split does not follow the artifact-strength ordering: DINOv2 B/14 has the highest ρ_{art} in the entire panel (57.9) yet is unaffected, while CLIP at 23.2 collapses. Artifact severity and functional importance are distinct properties. SigLIP has no CLS token, so its readout is necessarily pooled, exposing it to exactly the concern raised in Section 8. Holding the model, layer, and ablated tokens fixed, we varied only the pooling operator, including norm-mean, which L_2-normalises every token before averaging and therefore cannot be dominated by token magnitude (Table 5). The drop

Table 4: Global-readout k-NN accuracy under ablation of the 8 highest-norm patch tokens (196 tokens at patch size 16, 256 at patch size 14) ($N = 3000$, $1\sigma \approx 1.26$ points). Bootstrap 95% CIs in brackets; "sig." means the intervals do not overlap. Dose–response curve in Figure 17.

Model	Baseline	Ablate top-8	Sig.?
CLIP B/16 (Radford et al., 2021)	82.4 [80.2, 84.1]	67.9 [65.1, 69.8]	**yes**
SigLIP B/16 (Zhai et al., 2023)	84.2 [82.3, 85.8]	70.9 [69.1, 72.9]	**yes**
DINOv2 B/14	92.6 [91.3, 93.9]	91.0 [89.6, 92.5]	no
DINOv2 L/14	91.9 [90.5, 93.3]	92.5 [91.1, 93.9]	no
DINOv2 B/14 +reg	92.6 [91.2, 94.0]	92.5 [91.2, 94.0]	no
DeiT-III B/16 (Touvron et al., 2022)	89.7 [87.9, 91.3]	89.5 [87.9, 91.2]	no
DINO B/16 (Caron et al., 2021)	86.7 [84.9, 88.2]	86.3 [84.3, 87.9]	no

Table 5: Accuracy drop under top-8 ablation across pooling operators ($N = 1000$, 500-image gallery, $1\sigma \approx 2.19$). The effect survives magnitude-free pooling in both contrastive models and is absent in both controls.

Model	mean	norm-mean	median	trim-mean	random ctrl
SigLIP B/16	20.0	**21.2**	21.2	21.0	−1.0
CLIP B/16	18.2	**18.2**	15.2	18.4	−0.9
DINOv2 B/14	0.0	0.8	−2.6	0.8	2.0
DINO B/16	−0.4	−0.4	−3.6	1.0	0.0

is essentially invariant to the operator, with values of 20.0, 21.2, 21.2, and 21.0 for SigLIP and 18.2, 18.2, 15.2, 18.4 for CLIP, while the matched random controls remain within one point of zero in both models. The effect is therefore representational rather than an artifact of aggregation. Ablation makes representations of different images markedly more similar to one another. Mean pairwise cosine similarity rises by +0.210 for SigLIP and +0.092 for CLIP, compared with +0.013 for DINOv2 and +0.018 for DINO, while random ablation produces no such change (+0.007, +0.001). The artifact tokens therefore carry image-discriminative content. Removing them leaves a largely shared direction, causing nearest-neighbour retrieval to fail as images collapse toward a common representation. The concurrent rise in effective rank (+22.1, +17.1) is not contradictory. Effective rank measures how evenly variance is distributed across directions, so removing a dominant informative subspace can leave a more isotropic but less informative residual. Full geometry statistics are reported in Section I. The magnitude of the effect depends on the evaluation domain. SigLIP loses 20.0 points on Oxford-IIIT Pet but only 1.8 on EuroSAT under an identical intervention at an identical layer. Combined with Table 1, which shows SigLIP's artifact strength differing by only 23% between these datasets, this separates two easily conflated properties. Artifact tokens form identically regardless of input distribution, but what they carry depends on the task and training objective.

8 F5: The sign of the ablation effect is set by the training objective

Figure 1 showed the same intervention producing opposite-signed conclusions under two evaluation protocols. This section resolves that discrepancy. The sign is not a property of the protocol; it is a property of the model, and specifically of its training objective. The two protocols of Figure 1 differ along two axes at once, pooled versus per-token readout and distance-based versus learned evaluation. Figure 4a reports three of the four combinations on DINOv2 L/14; the fourth is the segmentation probe of Table 6. Per-token k-NN, which involves no pooling at all, shows the largest gain (+17.6 points), while the pooled linear probe shows almost none (+2.6). Pooling is therefore not the operative variable. Extending the same decomposition across models reverses the sign for CLIP on every evaluator (−17.2, −6.2, −13.3) and leaves DINO, which has no artifacts, flat throughout. The effect is stable across neighbourhood sizes from $k = 1$ to $k = 100$ (Figure 4b), ruling out a density artifact. The pooled–per-token axis therefore does not determine the sign. Dense segmentation (Table 6) is a separate case: it decreases for every model regardless of objective, because

Table 6: Linear segmentation probe on Oxford-IIIT Pet trimaps, frozen features. Segmentation quality decreases for every model, and high-norm ablation is indistinguishable from random ablation, in contrast to the k-NN results of Figure 5, where the two separate by 5–20×.

Model	mIoU base	mIoU top-8	mIoU rand-8	Δ	Border IoU (base → top-8)
DINOv2 B/14	75.18	73.45	73.36	−1.73	48.43 → 47.64
DINOv2 L/14	73.57	71.04	71.66	−2.53	45.05 → 42.44
DINOv2 B/14 +reg	75.60	72.88	73.47	−2.72	49.24 → 46.26
CLIP B/16	70.32	69.71	67.98	−0.61	39.96 → 39.69
DINO B/16	76.26	72.44	73.69	−3.81	50.74 → 45.49

per-pixel localisation depends on spatial content that ablation removes irrespective of what those tokens encode. Reporting a single ablation layer would leave open whether the result is an artifact of that choice, so we sweep the neighbourhood of each model's peak-artifact layer (Figure 5). Three findings follow. The sign is invariant within each model: DINOv2 is positive at every layer tested, CLIP is negative at every layer, and DINO is indistinguishable from its control throughout. The magnitude varies by a factor of two to three across the band, so the peak-artifact convention used in Table 3 and Table 4 is conservative. CLIP's largest effect is -33.3 at layer 5, compared with the -13.3 reported at layer 7. The effect also exceeds its matched random control by 5–20×. With random ablation averaged over five seeds, controls remain within ± 2 points with standard deviations of 0.1–3.9, while top-8 ablation reaches $+18.1$ and -33.3. Within DINOv2, the maximal effect grows from $+9.8$ points at 87M parameters to $+18.1$ at 304M, mirroring the read-score emergence of Section 5 and indicating that the same underlying phenomenon drives both. The number of tokens required to produce the full effect differs sharply across models (Section D). In DINOv2 L/14, ablating the single highest-norm token recovers $+17.6$ of the $+18.1$ maximum, and the per-image adaptive criterion selects an average of 1.0 token. In DINOv2 B/14, the effect accumulates over roughly five tokens, and in CLIP B/16 over six. The fact that one token out of 256, corresponding to 0.4% of the sequence, can move nearest-neighbour accuracy by 17 points places a strong constraint on any account of what these tokens do. We tested whether the change in representation geometry accounts for the direction of the effect. It does not. Under ablation, the effective rank of the token cloud falls in DINOv2 (-26.3 at L/14) and rises in CLIP ($+41.3$), which superficially tracks the accuracy split. However, this correspondence breaks within a single model. In SigLIP, the ablation layer producing the largest accuracy gain ($+6.0$ at layer 7) also produces a rank increase ($+10.8$), while the layer producing a rank decrease (-13.9 at layer 10) produces no gain ($+0.3$). Common-mode and pairwise-similarity statistics likewise rise under ablation in every model, including DINO, where nothing happens functionally. This indicates that these statistics partly track the mechanical effect of removing any eight tokens. Artifact tokens play opposite functional roles depending on the training objective. In image–text contrastive models, they carry image-discriminative content, and removing them degrades every evaluator we tested. In self-distillation models, they carry no such content but distort the representation, and removing them improves distance-based evaluation while leaving learned probes roughly unchanged. This unifies the two observations of Section 7 and Figure 1. The apparent protocol dependence arises because pooled k-NN and per-token probing were applied to models on opposite sides of this split, rather than because the protocols themselves disagree. We can state what determines the sign without yet explaining the underlying mechanism. Why a self-distillation objective produces high-norm tokens that are actively harmful to retrieval, while a contrastive objective produces high-norm tokens that carry the signal, remains open. Table 6 shows the practical consequence for dense evaluation: a per-token linear segmentation probe records a decrease for every model, so pooled retrieval metrics and per-token probes should not be treated as interchangeable when measuring artifact impact.

9 Limitations and Conclusion

Limitations. F5 establishes what determines the sign of the ablation effect but not why: we do not explain why a self-distillation objective produces high-norm tokens that harm retrieval while a contrastive objective produces tokens that carry the signal. Segmentation results use Oxford-IIIT Pet trimaps only; ADE20K or Pascal VOC would strengthen them, and the probe uses a fixed 40 epochs without early stopping, so it is valid for within-model comparison but not across architectures

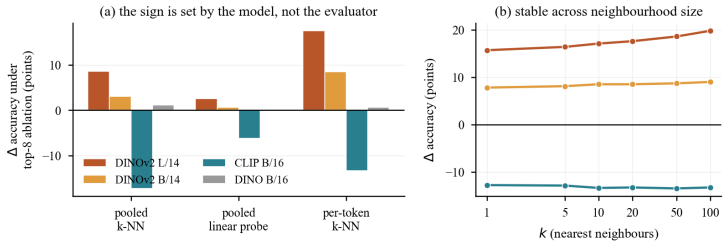


Figure 4: **(a)** Change in accuracy under top-8 ablation across three evaluators. The sign is constant within each model and differs between models: pooling is not the operative variable. **(b)** The effect is stable across neighbourhood sizes spanning two orders of magnitude.

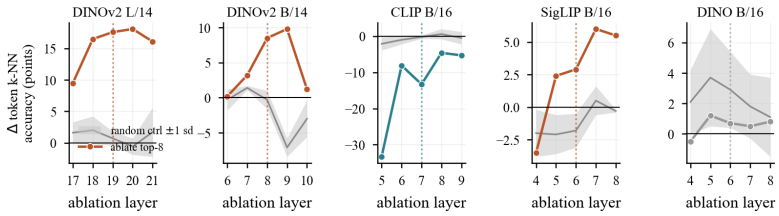


Figure 5: Change in per-token k-NN accuracy under ablation of the eight highest-norm patch tokens. Shaded band is the matched random-token control, averaged over five seeds, ± 1 standard deviation. DINOv2 is positive at every layer, CLIP negative at every layer, DINO within its control everywhere. The dotted line marks the peak-artifact layer, which is a conservative rather than maximal choice.

with different feature dimensionality. Attention statistics use 16–32 images, adequate for aggregate measures but not for per-image variance claims. We show artifact strength is input-invariant at inference but do not train models on different distributions to confirm that formation is determined during pretraining. Gallery size differs between Table 4 ($N = 3000$) and Table 5 ($N = 1000$), so only within-table comparisons are valid.

Conclusion. Artifact tokens in Vision Transformers are produced by model weights rather than by input statistics. Across nine visual domains and a $9.1\times$ range in low-information content, artifact strength varies by at most 10% and the layer at which artifacts peak does not move at all. The behaviour emerges with scale between 22M and 87M parameters, and register tokens suppress it by capturing the read target rather than by removing the underlying pressure. For image–text contrastive models these tokens carry the image-discriminative content: removing 4% of tokens costs 15–21 accuracy points under every pooling operator we tested. The sign of the ablation effect is determined by the training objective rather than by the evaluation protocol, invariant across layers, evaluators and neighbourhood sizes; representation geometry changes substantially under ablation but does not predict that sign. We also report a negative architectural result, per-edge attention gating offers no practical capacity advantage over a far cheaper additive gate (Section J).

What this enables. Three of our results change how subsequent work in this area should be conducted. Because artifact behaviour does not depend on the input distribution (Section 4), a single dataset suffices to characterise a model, and cross-dataset variation in reported artifact strength should be attributed to measurement protocol rather than to the data. Because pooled retrieval and per-token probing can disagree in sign (Figure 1), interventions on frozen features should be validated under both before a claim about representation quality is made. And because the direction of the ablation effect is set by the training objective (Section 8), a mitigation validated on a self-distillation backbone may be harmful on a contrastive one; proposed methods should be tested on both families. We release the diagnostic suite of scale-free artifact strength, read score, and matched-control ablation so that these measurements can be compared consistently across papers. Section K records six hypotheses that we tested and rejected, including two candidate mechanisms for Section 8, to help others avoid repeating these experiments. The question we leave open is now sharper than the one we started with. Why does a self-distillation objective produce high-norm tokens that harm retrieval, while a contrastive objective produces tokens that carry the signal?

USE OF AI TOOLS

The research idea, research direction, problem formulation, and overall scientific objectives were developed by the authors. The authors independently conducted the literature review and carefully studied the relevant prior work that forms the basis of this study in order to identify, understand, and formulate the research problem. All major research decisions, including the choice of methodology, experimental questions, and interpretation of the findings, were made by the authors. A large language model (Claude, Anthropic) was subsequently used as a research assistant to support methodology development, implementation, experimentation, literature discovery, and writing.

Methodology development, implementation, and experiments: Within the research direction and problem formulation established by the authors, the model assisted with the design and refinement of the methodology, including implementation details, candidate experimental protocols, and control experiments. It also helped design and implement experiments such as the matched random-ablation protocol, threshold and layer sweeps, and the evaluator decomposition reported in Section 8. The model wrote a substantial portion of the analysis and figure-generation code. All code was reviewed and executed by the authors, and the authors determined which experiments to conduct and which results to report. Every reported numerical result was obtained from experiments executed by the authors; no numerical result was generated, estimated, or fabricated by the model.

Literature discovery: The main papers underlying the study and the literature used to define and motivate the research problem were identified and reviewed independently by the authors. The model was used only as a supplementary research assistant for web-based literature searches and to surface potentially relevant additional works. All citations and relevant claims were independently checked and verified by the authors against the primary sources.

Writing: The model assisted with drafting, restructuring, and copy-editing portions of the manuscript. All text was reviewed and revised by the authors, who take responsibility for the final content.

REPRODUCIBILITY STATEMENT

All experiments use publicly available checkpoints from `timm` (Wightman, 2019) and publicly available datasets (Section E). No training was performed. Hardware: a single NVIDIA TITAN X (Pascal, 12 GB); software: PyTorch 2.5.1, `timm` 1.0.27, CUDA 11.8. All computations use fp32; on Pascal hardware `torch.cuda.is_bf16_supported()` returns `True` despite the absence of bf16 hardware, and we measured fp16 at $1.01\times$ and bf16 at $1.20\times$ the runtime of fp32. Thresholds, layer-selection rules and pooling operators are specified in Section 3. Code, configuration files and figure-generation scripts are in the supplementary material.

REFERENCES

Spiros Baxevanakis, Platon Karageorgis, Ioannis Dravilas, and Konrad Szewczyk. Do all vision transformers need registers? a cross-architectural reassessment. *arXiv preprint arXiv:2603.25803*, 2026.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. arXiv:2309.16588.

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.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Lukas Fesser, Mozes Jacobs, Thomas Fel, Andy Keller, and Sham Kakade. A unifying view of attention sinks: Two algorithms, two solutions. *arXiv preprint arXiv:2606.08105*, 2026.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Boris Iglewicz and David C Hoaglin. *Volume 16: how to detect and handle outliers*. Quality Press, 1993.

Nicholas Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. *Advances in neural information processing systems*, 38:56557–56595, 2026.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Ivan Titov, Chang Zhou, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, 2025. Oral presentation. arXiv:2505.06708.

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*, 2021.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *European Conference on Computer Vision*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Pavlo Vasylenko, Hugo Pitorro, Andre Martins, and Marcos V Treviso. Long-context generalization with sparse attention. In *International Conference on Learning Representations*, pp. 15256–15303, 2026.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. softmax is not enough (for sharp out-of-distribution). *arXiv preprint arXiv:2410.01104*, 2024.

Ross Wightman. PyTorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 – a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. In *International Conference on Learning Representations*, 2025. arXiv:2410.05258.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Appendix Overview

The appendix contains supporting material in the order the main text refers to it. Nothing here is required to follow the argument, but each section either substantiates a claim made in passing, reports a control we ran, or gives the full numbers behind a summary statistic. Section A provides additional background details that support the methodology. Section B provides method details and qualitative validation. It includes a threshold-sensitivity analysis verifying that our results do not depend on the outlier cutoff (Table 7), the observation that read score peaks two to three blocks deeper than token norm, and qualitative maps showing what the diagnostic quantities look like on real images (Figure 6–Figure 8). Section E provides the evaluation panel, including full citations, sample counts, and low-information statistics for the ten datasets of Table 1, together with representative images from each domain. Section F reports the full checkpoint sweep. It gives artifact strength and read score for all 15 checkpoints (Table 11), describes the two-cluster structure that justifies treating artifact severity as a single axis, and discusses the unexplained CLIP B/32 within-family anomaly. Section G presents the cross-domain supplementary analysis. It reports per-domain token-norm distributions and depth profiles, showing the layer invariance of Table 1 as continuous curves rather than as a column of identical integers. Section H presents the depth profile and attention handoff. It examines the sign change in the norm/variance correlation with depth and the layer-resolved handoff in which CLS serves as the attention target early and late, while patch outliers take over through the middle of the network. Section I provides supplementary results for F3–F5. It reports pooled and token-level representation geometry under ablation (Table 14, Table 13), including the caveat that common-mode and pairwise-similarity statistics rise in every model, even DINO, where the functional effect is null. These results therefore partly reflect the mechanical consequence of removing any eight tokens. Section J examines per-edge attention gating. It reports a negative result obtained without training: extending gated attention from per-query to per-query–key gating captures 89–99% of the available structure with an additive gate alone, leaving too little headroom to justify the additional cost. This addresses a question left open by Fesser et al. (2026). Finally, Section K records the hypotheses tested and rejected. It describes the six hypotheses evaluated against falsification criteria specified in advance and presents the deciding evidence for each, so that others need not repeat the work.

A Extended background

Why the sum-to-one constraint forces emission: Because the attention weights in Equation (1) are non-negative and sum to one, O_i lies in the convex hull of $\{v_j\}$ and cannot be the zero vector unless the value vectors themselves cancel. A head therefore has no mechanism for abstaining at a given position. The model's workaround is to drive some v_j toward an uninformative value and route surplus attention there, which requires designating a token to serve that role. This is the mechanism behind attention sinks in language models (Xiao et al., 2024), and it explains why the remedies proposed in that literature all relax the constraint: learnable sink logits add a term to the softmax denominator, output gating (Qiu et al., 2025) scales the emitted vector after the fact, and differential attention (Ye et al., 2025) subtracts two softmax maps so that effective weights can be zero or negative.

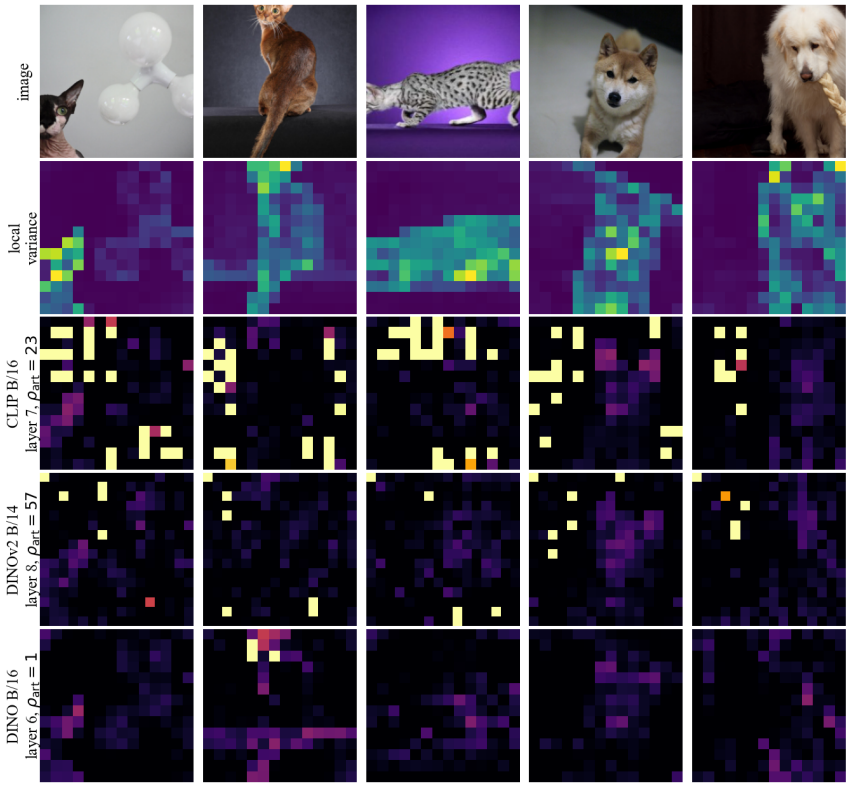


Figure 6: Rows: input images; local pixel variance per patch; robust z-scores of token norms at each model's peak-artifact layer, on a shared colour scale.

Dispersion at long context: Veličković et al. (2024) prove a related consequence of the same normalisation: with bounded logits, softmax coefficients necessarily disperse as sequence length grows, so a head cannot remain sharply focused at long context regardless of training. Vasylenko et al. (2026) recover bounded entropy by replacing softmax with α-entmax, which admits exactly-zero weights. We do not pursue these directions here, but they establish that the normalisation constraint has consequences beyond sink formation.

Why the vision case is harder to measure: Language models have a canonical first position that reliably serves as the sink, so attention mass on it can be read off a fixed index. Vision Transformers have no such position: the absorbing tokens are discovered from image content and differ between images and between layers. Measuring the analogous quantity therefore requires identifying the outlier set per image and per layer, which is what Equation (2) does and what distinguishes our read score from the sink-rate statistics reported in the language literature.

B Method details and qualitative validation

We use $z > 5$ rather than the conventional 3.5 because the effect sizes here are extreme. Table 7 reports the full sweep. Over $z \in [3, 20]$ — a $6.7\times$ range — read score changes by 0.008 for CLIP and 0.016 for DINOv2, and the peak layer does not move. The number of outliers per image falls from 8.5 to 4.5 (CLIP) and 6.7 to 1.4 (DINOv2), so the statistic is stable even as the outlier set shrinks by a factor of five. For comparability with Baxevanakis et al. (2026) we also computed the 98th-percentile convention; conclusions are unchanged.

We also note that read score peaks at layer 10 in both models, two to three blocks deeper than the peak-norm layers of 7 and 8 reported in Table 1. High-norm tokens therefore form before they are maximally read from, consistent with the handoff structure of Section H.

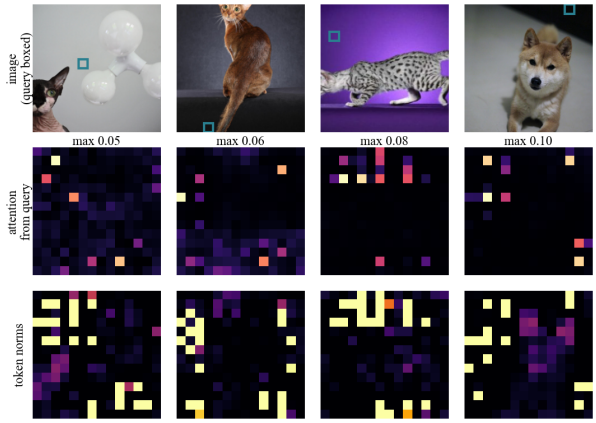


Figure 7: Where a background query sends its attention. Query patch boxed in the top row; attention distribution at CLIP B/16's peak layer in the middle row; token norms in the bottom row.

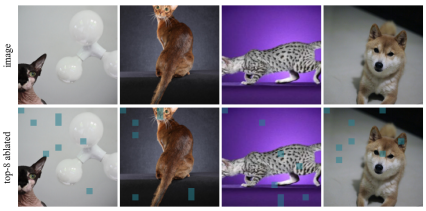


Figure 8: The ablation intervention. Highlighted patches are the eight highest-norm tokens at DINOv2 B/14's peak layer.

Table 7: Outlier-threshold sensitivity. Artifact strength is threshold-free by construction; read score and outlier count are not. Peak layer is the layer of maximal read score.

	CLIP B/16				DINOv2 B/14			
z	outliers	ρ_{art}	s_{read}	peak	outliers	ρ_{art}	s_{read}	peak
3.0	8.49	23.1	0.571	10	6.68	57.2	0.398	10
3.5	7.24	23.1	0.568	10	4.85	57.2	0.396	10
5.0	5.79	23.1	0.565	10	2.82	57.2	0.394	10
8.0	5.17	23.1	0.565	10	1.88	57.2	0.393	10
12.0	4.86	23.1	0.564	10	1.63	57.2	0.389	10
20.0	4.52	23.1	0.563	10	1.38	57.2	0.382	10

Figure 6 shows what the diagnostic quantities look like in practice. The robust z-maps for CLIP and DINOv2 contain a handful of bright tokens located in regions that are dark in the corresponding local-variance map, that is, in low-information patches. DINO, the artifact-free exception documented by Darcet et al. (2024), shows no such concentration, confirming that the procedure does not manufacture structure where none exists.

Figure 7 makes the read behaviour concrete at the level of a single query: for each image we select the patch with the lowest local variance and plot where its attention goes. Attention from uninformative regions concentrates on the high-norm tokens rather than on semantically related content. Figure 8 visualises the ablation intervention.

C Ruling out saturation

Section 4 shows that artifact strength is nearly constant across a $9.1\times$ range in low-information content. A plateau of this kind admits two readings. Under the first, input statistics genuinely do not modulate formation. Under the second, every dataset in the panel supplies a surplus of candidate sites, so the measurement is saturated and would fall only below some threshold not reached by any dataset we tested. The second reading is plausible: DTD, the lowest-flat dataset at 6.4%, still provides roughly twelve low-information patches per image, while Table 9 shows that a model needs between one and nine.

We separate the two readings with synthetic inputs constructed to contain no flat patches. Uniform pixel noise and low-pass-filtered noise both measure 0.000% under the criterion of Section 4, against 6.4% for DTD. The second input is included because uniform noise lies far outside any natural image distribution, so a reduction in artifact strength could reflect distribution shift rather than the absence

of flat patches; filtering restores natural-image-like local structure while keeping the flat fraction at zero.

Table 8: Artifact strength and peak layer on synthetic inputs containing no flat patches, against the panel means of Table 1.

Model	uniform noise ρ_{art}	uniform noise layer	smoothed noise ρ_{art}	smoothed noise layer	panel mean (layer)
CLIP B/16	19.1	7	19.5	7	23.8 (7)
DINOv2 B/14	49.5	8	51.7	8	57.6 (8)
SigLIP B/16	20.7	5	20.8	5	21.9 (6)

Table 8 reports the result. Artifacts form at 80–95% of panel-mean strength with no flat patches available. The peak layer is unchanged for CLIP and DINOv2 and moves by one block for SigLIP. The two synthetic inputs agree to within 2.2, so the result is not driven by the off-distribution character of uniform noise.

A saturation account predicts collapse at zero availability, not a reduction of 5–20%. Formation therefore does not require low-information patches to be present; such regions determine where artifacts land when they exist, and the mild reduction in strength is consistent with placement being less efficient when no natural destination is available. We note that the synthetic inputs lie outside any natural image distribution, so the residual reduction cannot be cleanly attributed to the absence of flat patches rather than to distribution shift; the invariance of the peak layer is the stronger part of this result.

D Sensitivity to the number of ablated tokens

The ablation count $k = 8$ used throughout the main text is fixed independently of the $z > 5$ detection criterion of Equation (2), so that the random-ablation control is exactly matched in size. Table 9 verifies that this choice is not load-bearing.

Table 9: Change in per-token k-NN accuracy under ablation of the k highest-norm patch tokens, at each model's peak-artifact layer. The adaptive column ablates exactly the tokens satisfying $z > 5$ for each image, with the mean number (n) removed in parentheses. Sign and approximate magnitude are stable across an order of magnitude in k.

Model	k=1	2	3	4	5	6	8	12	16	adaptive
DINOv2 L/14	+17.6	+18.0	+17.9	+18.0	+17.9	+17.7	+17.6	+17.2	+17.1	+18.2 (n=1.0)
DINOv2 B/14	+0.2	+3.0	+6.4	+8.0	+8.4	+8.7	+8.5	+7.8	+7.3	+8.9 (n=5.0)
CLIP B/16	−0.1	−0.6	−2.4	−6.3	−11.2	−13.4	−13.3	−12.8	−13.9	−11.0 (n=9.2)
DINO B/16	+0.3	+0.7	+0.8	+0.8	+0.8	+0.8	+0.7	+1.0	+1.0	−0.0 (n=0.6)

Three observations: First, the adaptive column agrees with fixed $k = 8$ to within 2.3 points in every model, so tying the ablation count to the per-image outlier threshold changes nothing material. Second, the sign is stable from $k = 1$ upward in every model, so no conclusion in the paper depends on the choice. Third, the number of tokens required to reach the full effect differs sharply between models and is informative in itself: DINOv2 L/14 saturates at $k = 1$, DINOv2 B/14 at $k \approx 5$, and CLIP B/16 at $k \approx 6$, while DINO never departs from its control. The adaptive counts track this ordering (1.0, 5.0, 9.2 and 0.6 tokens respectively), indicating that the detection criterion and the functional effect identify the same token population. The $k = 8$ convention is therefore conservative for DINOv2 L/14, where a single token carries the entire effect, and approximately optimal for CLIP.

E Evaluation panel

Table 10 gives the full dataset panel with citations; Figure 9 shows representative samples. The panel was assembled to maximise the range of the independent variable, low-information content, while spanning training distributions the evaluated models never saw. We required datasets that were

publicly available, loadable without manual preprocessing, and large enough to draw 800 images without exhausting a split. Within those constraints we selected for domain diversity rather than for task difficulty: the artifact measurements of Table 1 depend only on the images, not on their labels, so class structure is irrelevant to F1.

For each image we compute the per-channel standard deviation within every non-overlapping 16×16 patch, average across channels, and report the fraction of patches falling below 0.02 on the $[0, 1]$ intensity scale. This threshold was chosen so that visually uniform regions, clear sky, plain studio backgrounds, the black border of a chest radiograph, are captured while lightly textured regions are not. Table 10 gives the resulting statistics; the panel spans 6.4% (DTD) to 58.6% (EuroSAT), a $9.1\times$ ratio or 52 percentage points.

Four of the ten datasets fall outside the natural image distribution these models were pretrained on. PathMNIST (haematoxylin and eosin stained tissue) and BloodMNIST (stained blood smears) are microscopy; ChestMNIST is radiography; DermaMNIST is clinical dermoscopy. These matter because they represent a distribution shift large enough to perturb any input-driven mechanism, and because they occupy the middle and upper portions of the low-information range. We verified that the MedMNIST v2 images are natively 224×224 rather than upsampled from the 28×28 variants: the ratio of fine-scale (4×4) to coarse-scale (16×16) patch variance falls between 0.35 and 0.57 for all four, consistent with genuine high-resolution imagery. An upsampled image would show a ratio below 0.3, because interpolation removes fine detail while leaving coarse structure intact.

All images are resized so the shorter side is 256 pixels, centre-cropped to 224×224, and converted to RGB (the MedMNIST grayscale sets are replicated across channels). Model-specific normalisation is applied afterwards using each checkpoint's own statistics from `timm`. Because normalisation differs between models, the low-information statistics in Table 10 are computed on the unnormalised $[0, 1]$ images so that a single value characterises each dataset independently of which model consumes it. We draw 800 images per dataset with a fixed seed; the identical image set is used for every model, so cross-model comparisons in Table 1 are exactly paired.

Table 10: Evaluation panel, ordered by low-information content. $n = 800$ images per dataset, resized to 224×224.

Dataset	Domain	% flat patches	Mean patch variance
DTD (Cimpoi et al., 2014)	texture	6.4	0.1082
Flowers102 (Nilsback & Zisserman, 2008)	natural	9.3	0.0955
Food-101 (Bossard et al., 2014)	dense	10.5	0.0921
Oxford-IIIT Pet (Parkhi et al., 2012)	natural	12.3	0.0873
PathMNIST (Yang et al., 2023)	histopath.	16.1	0.0995
ChestMNIST (Yang et al., 2023)	radiography	19.5	0.0618
BloodMNIST (Yang et al., 2023)	microscopy	33.2	0.0526
FGVC-Aircraft (Maji et al., 2013)	sky bg.	41.0	0.0739
DermaMNIST (Yang et al., 2023)	dermatology	45.0	0.0299
EuroSAT (Helber et al., 2019)	satellite	58.6	0.0264

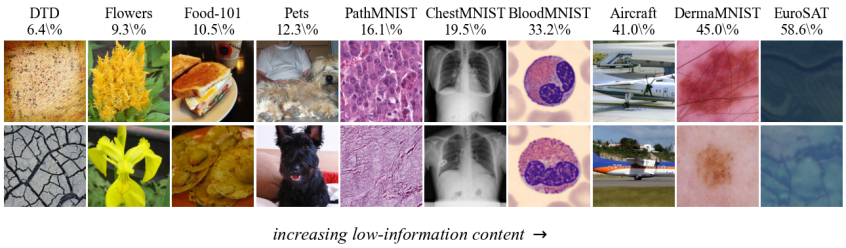


Figure 9: Two samples from each dataset, ordered by low-information content.

F FULL CHECKPOINT SWEEP

Table 11 reports artifact strength and read score for all 15 checkpoints, and Figure 10 plots the same data. Figure 11 shows the scale emergence of Table 2: the base models rise sharply between 22M and 87M parameters while the register-trained variants remain flat near 0.01 at every scale. The five models above the rule all exceed $s_{\text{read}} = 0.38$ and $\rho_{\text{art}} = 9.2$, while every model below falls under $s_{\text{read}} = 0.12$ and $\rho_{\text{art}} = 2.5$. No checkpoint occupies the intermediate region, which justifies treating artifact severity as a single axis.

Table 11: Full sweep. s_{read} is attention mass received by outlier tokens at the peak layer.

Model	Family	ρ_{art}	s_{read}
vit_base_patch16_siglip_224	SigLIP	19.7	0.628
vit_base_patch16_clip_224	CLIP	23.2	0.575
vit_large_patch14_clip_224	CLIP	29.4	0.468
vit_base_patch16_224.augreg	Supervised	9.2	0.449
vit_base_patch14_dinov2	DINOv2	57.9	0.382
deit3_base_patch16_224	DeiT-III	2.4	0.120
deit3_small_patch16_224	DeiT-III	2.5	0.074
deit3_large_patch16_224	DeiT-III	2.0	0.030
vit_small_patch14_reg4_dinov2	DINOv2+reg	1.9	0.010
vit_base_patch16_224.dino	DINO	1.4	0.010
vit_base_patch14_reg4_dinov2	DINOv2+reg	1.6	0.007
vit_small_patch16_224.dino	DINO	1.6	0.007
vit_small_patch14_dinov2	DINOv2	1.9	0.007
vit_base_patch32_clip_224	CLIP	1.4	0.005
vit_base_patch16_224.mae	MAE	1.3	0.001

The CLIP B/32 anomaly: CLIP B/32 shows no artifacts ($\rho_{\text{art}} = 1.4$) while CLIP B/16 shows severe ones ($\rho_{\text{art}} = 23.2$), despite identical parameter count and training source. This is not a token-count effect: DINO B/8 (784 tokens) and DINO B/16 (196 tokens) both show $\rho_{\text{art}} \approx 1.5$. We report the within-family split as an open observation and as a second instance, alongside the DINO exception, of two closely matched models diverging sharply.

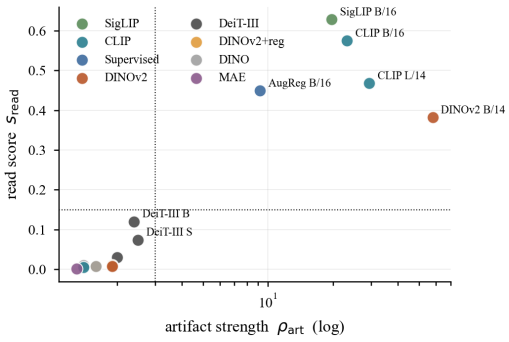


Figure 10: Artifact strength versus read score across 15 checkpoints, plotting Table 11. Two clusters, no intermediate cases.

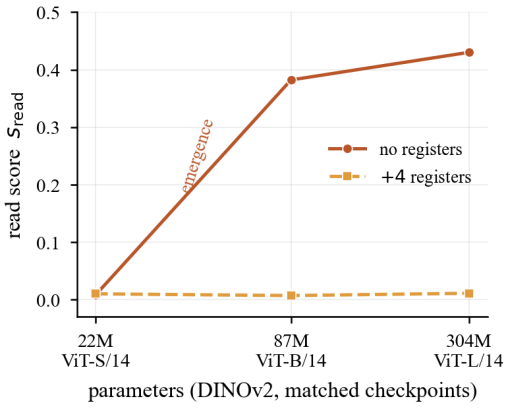


Figure 11: Read behaviour emerges with scale, plotting Table 2. Register-trained variants remain flat at every scale.

G CROSS-DOMAIN SUPPLEMENTARY ANALYSIS

Figure 12 plots the token-norm distribution at each domain's peak-artifact layer. Every distribution has the same shape, with a dense bulk and a sparse heavy tail extending well beyond three times

the median, regardless of whether the input is tissue, sky, or texture. Figure 13 overlays the artifact-strength-versus-depth profile for all domains, one panel per model. For the three strong-artifact models the curves are near-coincident and share a single maximum, which is the layer invariance of Table 1 rendered as a picture.

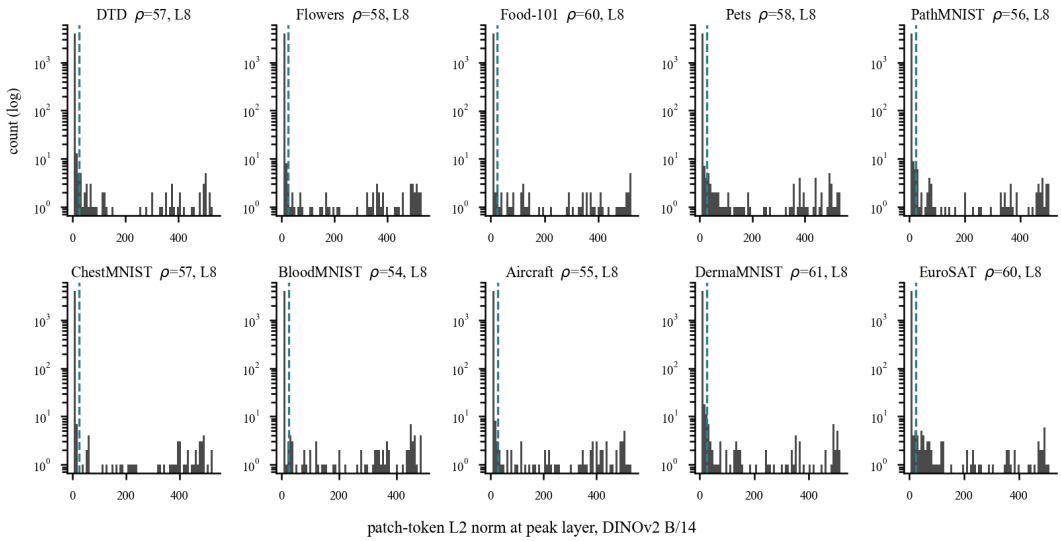


Figure 12: Token-norm distributions per domain, DINOv2 B/14, log-scaled counts. Dashed line marks three times the layer median.

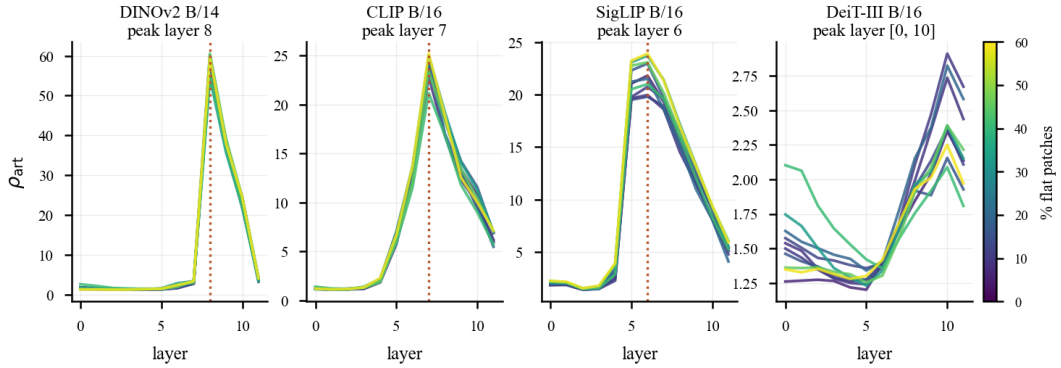


Figure 13: Artifact strength versus depth, all domains overlaid, coloured by low-information content. Dotted line marks the shared maximum.

H Depth profile and attention handoff

The Spearman correlation between token norm and local patch variance changes sign with depth (Figure 14). In DeiT-III S/16 it falls monotonically from $+0.849$ at layer 0 to -0.255 at layer 11, crossing zero at layer 7. CLIP B/16 crosses at layer 1; DINO never crosses. Early layers encode local content, so textured patches carry larger norms; beyond the crossing point, large norms migrate to flat regions.

In CLIP B/16 the destination of peak attention mass changes systematically with depth (Figure 15). At layers 0–3, 100% of queries direct peak attention to CLS. At layer 6 a handoff occurs (19% CLS, 26% outlier patches), and at layers 7–10 outlier patches receive peak attention from 88–97% of queries, with total mass on outliers rising to 0.572. At layer 11 the pattern reverses, with 97% returning to CLS. This is consistent with patch tokens serving as auxiliary storage through the middle

of the network and being released before readout. The shaded band is the dumping band referred to in Section J.

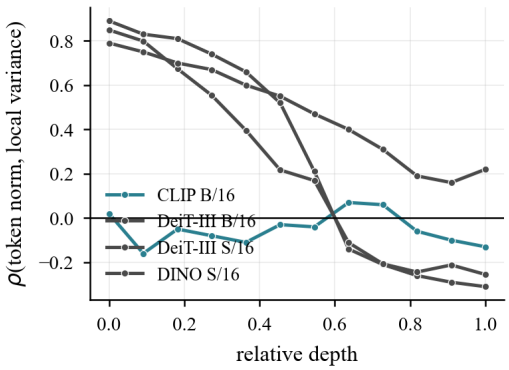


Figure 14: Spearman correlation between token norm and local patch variance versus depth. DINO never crosses zero.

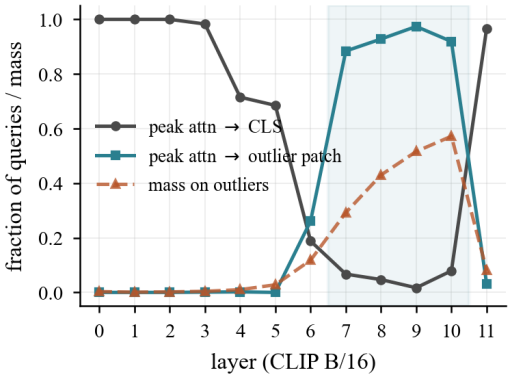


Figure 15: Destination of peak attention mass across depth in CLIP B/16.

I Supplementary results for F3, F4 and F5

Figure 16 plots the ablation damage of Table 3. Figure 17 gives the dose–response curve extending Table 4. Table 14 and Figure 18 give the pooled geometry statistics summarised in prose in Section 7; Table 13 gives the corresponding token-level statistics discussed in Section 8. Note that common-mode and pairwise similarity rise under ablation in every model, including DINO where the functional effect is null, so those two statistics partly reflect the mechanical consequence of removing any eight tokens rather than anything specific to artifact tokens. We quantify how evenly variance is distributed across directions using the participation ratio of the singular values, $\left(\sum_i \sigma_i^2\right)^2 / \sum_i \sigma_i^4$, computed on mean-centred features. This differs from the entropy-based effective rank of Roy & Vetterli (2007). We use the participation ratio because it is bounded by the ambient dimension and remains stable under the small sample sizes used here. Only effective rank separates the two model families by sign, and Section 8 shows that it does not predict the sign of the accuracy change within a model. Figure 19 plots Table 6, and Figure 20 the norm-compression effect used to rule out the magnitude-domination explanation in Section 8.

Table 12 decomposes total attention mass at the peak-artifact layer into four destinations. Adding registers leaves the ordinary patch tokens almost untouched and instead transfers mass from CLS to the registers.

Table 12: Attention mass by destination, DINOv2 B/14 at its peak-artifact layer. Rows sum to one.

Model	CLS	registers	outlier patches	ordinary patches
DINOv2 B/14	0.319	—	0.018	0.663
DINOv2 B/14 +reg	0.052	0.304	0.003	0.641

This clarifies the CLS column of Table 3. Zeroing CLS in the register-trained model drives the representation to anti-correlation (-0.307), which is consistent with CLS having been relieved of its sink role and repurposed: it carries less surplus attention and correspondingly more task-relevant content.

J Per-edge attention gating offers no capacity advantage

Recent work gates attention outputs per query (Qiu et al., 2025). A natural extension gates per query–key pair, which would in principle allow a head to suppress one high-norm token while retaining another, an attractive prospect given the read behaviour of Figure 7. We test whether this

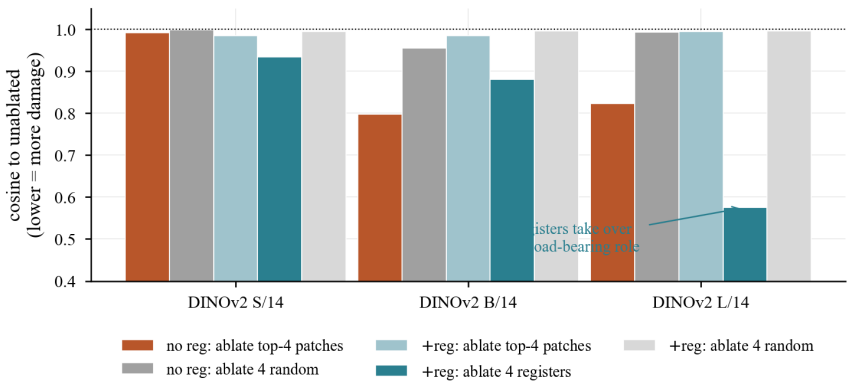


Figure 16: Registers capture the read target, plotting Table 3. Lower means more damage.

Table 13: Token-level representation geometry under top-8 ablation, at each model's peak-artifact layer. Δ rows give the change relative to no ablation. Accuracy is per-token k-NN with gallery/query split by image.

Model	Condition	token k-NN	common-mode	pairwise sim.	eff. rank
DINOv2 L/14	none	37.1	0.286	0.082	124.1
	top-8	54.6	0.381	0.145	97.8
	rand-8	39.9	0.292	0.085	119.4
	Δ	+17.6	+0.095	+0.064	−26.3
DINOv2 B/14	none	40.4	0.282	0.080	87.2
	top-8	48.9	0.346	0.120	71.8
	rand-8	42.7	0.289	0.083	86.6
	Δ	+8.5	+0.064	+0.040	−15.5
CLIP B/16	none	59.5	0.620	0.384	43.2
	top-8	46.2	0.765	0.586	84.5
	rand-8	60.2	0.617	0.380	44.7
	Δ	−13.3	+0.146	+0.202	+41.3
SigLIP B/16	none	44.0	0.464	0.215	56.4
	top-8	46.9	0.792	0.627	70.5
	rand-8	39.5	0.464	0.215	57.5
	Δ	+2.9	+0.328	+0.412	+14.1
DINO B/16	none	37.9	0.454	0.206	178.6
	top-8	38.6	0.470	0.221	179.3
	rand-8	45.8	0.463	0.215	175.8
	Δ	+0.7	+0.016	+0.015	+0.7

offers additional expressive capacity, without training, by decomposing attention maps from existing checkpoints.

Restrict the log-attention matrix to outlier columns and decompose it into row, column and interaction components, $M_{ij} = \mu + r_i + c_j + \varepsilon_{ij}$. A per-query gate can express only the row term, a per-key gate only the column term; the interaction term requires a genuinely pairwise gate. The variance explained by each family bounds what that family can achieve, independently of optimisation.

Table 15 reports variance explained per head on CLIP B/16 in the dumping band (layers 7–10, identified in Figure 15); Figure 21 plots the same data. The per-query column shows that the form used in prior work explains only 0.126–0.215 of the structure, so gating on the query alone is genuinely insufficient. Adding a per-key term closes almost the entire gap, reaching 0.889–0.994. The remaining headroom for a full pairwise gate is between 0.6% and 11%.

We further verified that head-averaging is not concealing per-head pairwise structure: mean pairwise agreement between heads on which outliers they target is 0.994–1.000. CLIP and DeiT-III require

Table 14: Change in representation geometry under top-8 ablation. Positive Δ pairwise similarity means representations of different images become less distinguishable.

Model	Δ common-mode	Δ pairwise sim.	Δ effective rank
SigLIP B/16	+0.120	+0.210	+22.1
CLIP B/16	+0.051	+0.092	+17.1
DINOv2 B/14	+0.010	+0.013	−6.6
DINO B/16	+0.010	+0.018	+1.8

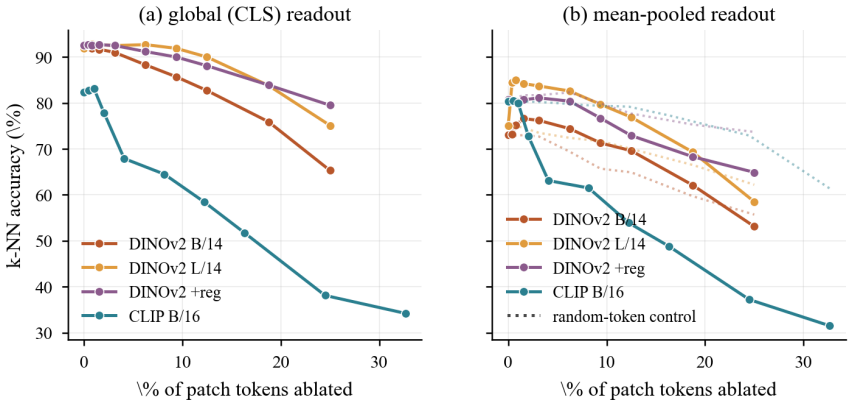


Figure 17: Dose–response under ablation. **(a)** Global (CLS) readout. **(b)** Mean-pooled readout, random-token controls dotted. The apparent DINOv2 L/14 gain in panel (b) is the effect examined in Section 8.

opposite gate families. Per-key gating dominates in CLIP (0.78 vs. 0.16), whereas per-query gating dominates in DeiT-III (0.82 vs. 0.12). No single pairwise design therefore serves both models.

In a preliminary implementation, per-edge gating cost 1.38× the runtime of standard attention and 4.1% additional parameters for that capacity gain. We find no justification for it. This addresses a question left open by Fesser et al. (2026), obtained without training any model.

Table 15: Variance explained by gate family, per head, CLIP B/16. An additive query+key gate captures 89–99% of what a rank-16 pairwise gate captures.

Layer	per-query	per-key	query+key	rank-2	rank-16
7	0.161	0.728	0.889	0.958	1.000
8	0.126	0.784	0.911	0.972	1.000
9	0.183	0.808	0.991	0.999	1.000
10	0.215	0.779	0.994	1.000	1.000

K Hypotheses tested and rejected

Negative results are rarely reported, but they consume real effort and their absence from the literature means the same dead ends get explored repeatedly. We therefore record in Table 16 the hypotheses we tested and rejected during this study, together with the experiment that decided each and a pointer to the evidence. Each was evaluated against a falsification criterion fixed before the experiment was run. In two cases, the two-mechanism taxonomy and the claim that ablation improves dense features, the hypothesis initially appeared to produce a positive result under the first evaluation protocol. It was rejected only after adding a control or a second evaluation protocol. Both are instances of the methodological point made in Section 8: an intervention on frozen features can appear to succeed under one reasonable evaluation and fail under another, and the discrepancy is visible only if both are run. The two rejected mechanisms for Section 8 are worth noting for anyone attempting to explain that result. Magnitude domination fails because the final LayerNorm compresses the norm ratio from

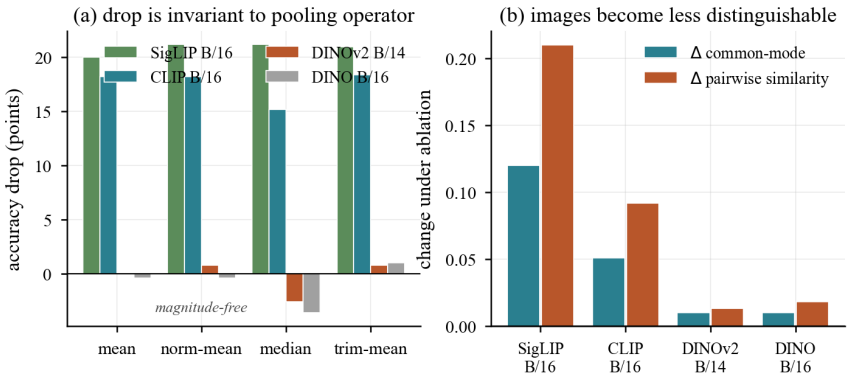


Figure 18: **(a)** Accuracy drop from Table 5 with only the pooling operator varied. **(b)** Geometry changes from Table 14.

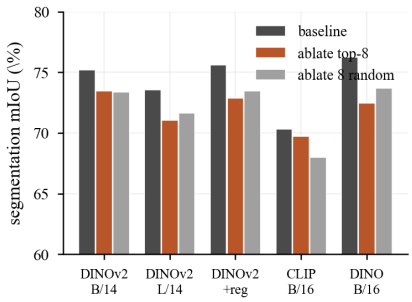


Figure 19: Linear segmentation probe, plotting Table 6.

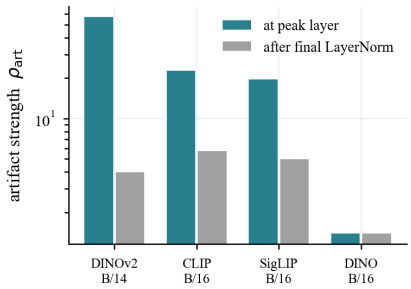


Figure 20: Artifact strength at the peak layer versus after the final LayerNorm (log scale).

57× at the artifact layer to 4× at the output, so the pooled features contain no magnitude outliers. Common-mode removal fails because ablation *raises* rather than lowers inter-image similarity, and does so in every model including those where the functional effect is null. A successful account will need to explain why the sign depends on the training objective while representation geometry does not track it.

Table 16: Rejected hypotheses, with the test applied and the outcome.

Hypothesis	Test	Outcome
Per-edge gating adds capacity over per-query gating	ANOVA decomposition of attention structure	Rejected: additive gate captures 89–99% (Table 15)
Artifacts fall into two mechanisms (read vs. write)	Read score across 15 checkpoints	Rejected: one severity axis, no intermediate cluster (Table 11)
Ablating artifacts improves dense features	Per-token linear segmentation probe	Rejected: mIoU decreases (Table 6)
Low-information input drives artifact formation	10 datasets, 9 domains, 9.1× range	Rejected: CV ≤ 0.102 (Table 1)
Pooled-vs-probe discrepancy is magnitude domination	Norm mass at readout layer	Rejected: LayerNorm compresses the norm ratio $57\times \to 4\times$ (Section K)
Pooled-vs-probe discrepancy is common-mode removal	Representation geometry under ablation	Rejected: $\Delta \leq 0.013$ (Table 14)

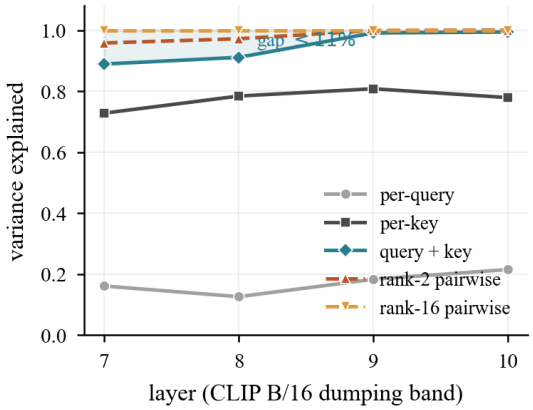


Figure 21: Per-edge gating adds little capacity. Shaded region is the gap between an additive gate and a rank-16 pairwise gate.