SAGA: SPATIALLY-AWARE GATED ATTENTION FOR SEPARATING OUTLIER STRUCTURE FROM FUNCTIONAL EFFECT IN VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers can develop a small number of unusually high-norm patch tokens, motivating methods that suppress, relocate or regularize this activity. Yet reducing an outlier count or producing a cleaner spatial map does not necessarily imply that the representation used by a downstream task has improved. We introduce **SAGA** (Spatially-Aware Gated Attention), a lightweight receiver-side mechanism that modulates patch-token attention outputs by layer, head and spatial position without adding extra tokens. Across three audited 300-epoch ImageNet-1K architecture-recipe cells with matched Baseline-SAGA checkpoints, SAGA reduces fixed-threshold high-norm-token counts by approximately 46–65% while yielding 0.327–0.456 percentage-point higher observed mean top-1 accuracy than the ViT baselines. Mechanistically, high-norm positions receive 6.15–$13.75\times$ more incoming attention per token than other patch positions and form highly reproducible spatial patterns. Matched frozen interventions, however, show that this spatial reproducibility does not translate into a reproducible coordinate-specific classifier response in the tested blocks. Functional relevance also depends on the readout: terminal patch-gate edits can alter patch representations while leaving CLS logits exactly unchanged, whereas, in a descriptive patch-level analysis, high-norm positions are associated with substantially poorer exact-transform correspondence. These results separate fixed-threshold outlier prevalence, spatial organization and readout-specific functional effect, motivating evaluation of ViT diagnostics against the downstream computation that consumes them.

1 INTRODUCTION

Vision Transformers (ViTs) process images as patch-token sequences, making patch representations central to computation and downstream use (Dosovitskiy et al., 2020). Some ViTs develop unusually high-norm patch tokens, often in low-information regions, that retain less local information while carrying more global information (Darcet et al., 2023). Register tokens can absorb this behavior and smooth feature and attention maps (Darcet et al., 2023), motivating efforts to suppress or relocate it. But if an intervention removes these outliers or cleans their spatial pattern, has the representation improved or only the diagnostic changed?

High-norm-token prevalence, spatial organization and downstream functional effect are distinct quantities: fewer flagged patches or smoother maps need not imply an improved downstream computation. Prior work shows that high-norm activity is spatially non-uniform, can be absorbed or redirected and that its removal can affect accuracy in some settings (Darcet et al., 2023; Jiang et al., 2025). What remains unclear is whether prevalence, spatial structure and readout-specific functional effect agree under controlled interventions. Without this separation, cleaner diagnostics can be mistaken for better representations. We therefore ask which properties of high-norm patch activity actually track downstream behavior and whether they can be controlled without adding tokens.

To address this gap, we introduce **SAGA** (Spatially-Aware Gated Attention), which modulates the attention update received by each patch through a learned spatial gate without adding tokens. Across three audited 300-epoch ImageNet-1K architecture–recipe cells with matched Baseline-SAGA checkpoints (Deng et al., 2009), SAGA yields fewer fixed-threshold high-norm tokens and higher

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.

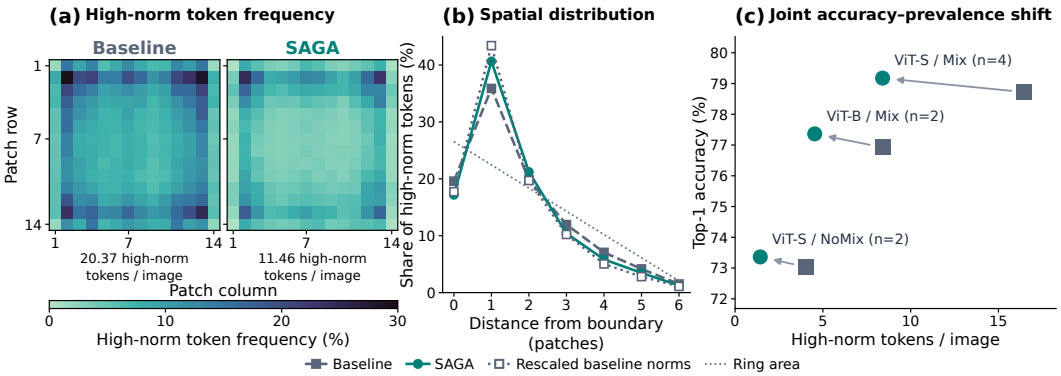

Figure 1: **SAGA reduces high-norm-token prevalence while improving observed ImageNet accuracy, even as spatial structure persists. (a)** Fixed-threshold high-norm-token frequency maps for one fresh ViT-S/16 mixing Baseline-SAGA pair; cell means are reported in Table 1. **(b)** Conditional high-norm-token distribution by boundary distance; the rescaled-baseline control globally scales baseline patch norms to match SAGA's token count, while ring area gives the expected share under uniform spatial occupancy. **(c)** 300-epoch ImageNet-1K cell means; all Baseline→SAGA trajectories move toward fewer high-norm tokens and higher observed mean top-1 accuracy. n denotes the number of paired Baseline-SAGA checkpoints.

observed mean top-1 accuracy than the corresponding ViT baselines (Fig. 1c; Table 1). Beyond ImageNet, supporting downstream evaluations show gains on both evaluated FGVC-Aircraft (Maji et al., 2013) backbones and COCO (Lin et al., 2014) overall AP, while multi-scale ADE20K (Zhou et al., 2017) performance is essentially unchanged.

Beyond its trained-system gains, SAGA exposes explicit spatial control, letting us test whether particular patch locations are functionally privileged. High-norm tokens are genuine attention sinks in our cohort, receiving 6.15–$13.75\times$ more incoming attention per token and their spatial organization is highly reproducible under the mixing recipe. Yet a simple global rescaling of baseline patch norms reproduces much of the preserved Baseline-SAGA profile. Matched frozen interventions further separate spatial reproducibility from classifier sensitivity: reproducible structure does not translate into a reproducible coordinate-specific response.

The diagnostic's functional meaning also depends on its consumer. As an exact architectural readout control, changing the terminal patch-only gate alters patch representations while leaving CLS logits unchanged across all eight SAGA checkpoints. In a separate descriptive patch-level analysis, high-norm positions are associated with substantially poorer exact-transform correspondence. Thus, the same diagnostic can be invisible to a CLS-only classifier while remaining informative about a patch-level readout. Lower high-norm-token counts or cleaner maps should therefore not be treated as evidence of better representations without testing the downstream computation that consumes them.

We make four contributions:

- **SAGA.** A lightweight, token-free receiver-side gate with layer-, head- and position-specific spatial modulation.
- **Trained-system gains.** Across all three available 300-epoch ImageNet-1K cells, SAGA yields fewer high-norm tokens and higher observed mean top-1 accuracy than the corresponding ViT baselines, with supporting gains on both evaluated FGVC-Aircraft settings and COCO overall AP while maintaining multi-scale ADE20K performance.
- **Mechanistic characterization.** Using attention-receiver measurements and a scale-matched control, we show that high-norm tokens are reproducible attention sinks, while much of their preserved spatial profile is reproduced by simple global norm rescaling.
- **Readout-aware analysis.** Using matched frozen interventions, we show that spatial reproducibility does not imply coordinate-specific classifier sensitivity and that functional relevance depends on the downstream consumer.

2 Related Work

High-norm patch artifacts in ViTs have prompted several explanations and mitigation strategies. Darcet et al. (2023) showed that high-norm patch tokens lose local information, can be absorbed by learned register tokens and exhibit non-uniform, border-biased spatial distributions. Wang et al. (2024) instead traced artifacts in DINOv2 (Oquab et al., 2024) to dominant singular structure in pretrained weights and proposed lightweight regularized fine-tuning. Jiang et al. (2025) identified sparse register neurons whose activity can be redirected to an added test-time token, while Chen et al. (2025) introduced post-hoc registers through self-distillation. More recently, Shi et al. (2026) attributed broader ViT artifacts to "lazy aggregation" through semantically irrelevant background patches and mitigated them with selective patch-to-CLS integration.

ViT representation artifacts extend beyond high-norm outliers. Yang et al. (2024) identified grid-like feature artifacts caused by positional embeddings and proposed cross-view denoising to improve dense prediction. Related modulation methods also differ from SAGA. Bondarenko et al. (2023) use gated attention to suppress activation outliers, while Qiu et al. (2025) study head-specific sigmoid gating after scaled dot-product attention in language models, showing that such gating can mitigate massive activations and attention-sink behavior. Their gates are query-dependent, whereas SAGA uses an input-independent gate indexed jointly by head and spatial position. LayerScale (Touvron et al., 2021b) instead applies learnable per-channel scaling to residual-block outputs. Prior work has also evaluated local information directly; Darcet et al. (2023) probed patch tokens through position prediction and pixel reconstruction, while Amir et al. (2021) used frozen ViT features as dense visual descriptors for correspondence. Taken together, prior work establishes that ViT artifacts can be spatially structured, mitigated, and associated with changes in local representations, but it remains unclear whether reducing their prevalence or changing their spatial organization predicts functional importance for the downstream readout.

3 Method and Measurement

3.1 SAGA: Spatially-Aware Gated Attention

SAGA is designed to control *where* attention updates are received without adding tokens or modifying the attention operator itself. It introduces a learned, input-independent gate indexed by transformer block, attention head and patch position, applied to patch receivers after scaled dot-product attention and before head reassembly and output projection. The principal ImageNet experiments start from randomly initialized ViT backbones rather than pretrained checkpoints.

Let $X_{\ell-1} \in \mathbb{R}^{(1+N)\times d}$ be the input to block $\ell \in \{1,\dots,L\}$, with one CLS token and N patch tokens. For head $h \in \{1,\dots,H\}$ of dimension d_h, let $Q_{\ell h}$, $K_{\ell h}$ and $V_{\ell h}$ denote the standard query, key and value projections of $X_{\ell-1}$. Scaled dot-product attention (Vaswani et al., 2017) gives

$$A_{\ell h} = \mathrm{softmax}\left(\frac{Q_{\ell h}K_{\ell h}^{\top}}{\sqrt{d_h}}\right), \qquad Z_{\ell h} = A_{\ell h}V_{\ell h}, \tag{1}$$

where $Z_{\ell h,p} \in \mathbb{R}^{d_h}$ is the attention output received at token position p before head concatenation and output projection. For patch positions $\mathcal{P} = \{1,\dots,N\}$, SAGA applies

$$g_{\ell hp} = \sigma(\phi_{\ell hp}), \qquad \widetilde{Z}_{\ell h,p} = g_{\ell hp}Z_{\ell h,p}, \qquad U_\ell = \mathrm{Concat}_{h=1}^{H}(\widetilde{Z}_{\ell h})W_\ell^O, \tag{2}$$

where $\phi_{\ell hp}$ is a learned gate logit, σ is the logistic sigmoid, W_ℓ^O is the standard output projection and U_ℓ is the attention-branch update before residual addition. The CLS row bypasses the gate, $\widetilde{Z}_{\ell h,\mathrm{CLS}} = Z_{\ell h,\mathrm{CLS}}$ and all other block operations remain unchanged (Fig. 2a).

The gate tensor $\Phi \in \mathbb{R}^{L\times H\times N}$ depends on block, head and spatial position, but not on image. For a fixed block input, SAGA leaves $A_{\ell h}$ unchanged and modulates only the attention contribution received by each patch. It adds no sequence tokens and only 14,112 parameters ($\approx 0.064\%$) to ViT-S/16 and 28,224 ($\approx 0.033\%$) to ViT-B/16 at 224×224 resolution ($N = 196$). All headline checkpoints initialize $\phi_{\ell hp} = 0$, so $g_{\ell hp} = 0.5$ and optimize the gates jointly with the backbone. Because this initialization also halves the initial patch attention update, Appendix G reports a separate single-seed 100-epoch control family with a frozen 0.5 gate, a learned per-head scalar, and a near-open spatial initialization ($\phi = 4$, $\sigma(\phi) \approx 0.982$).

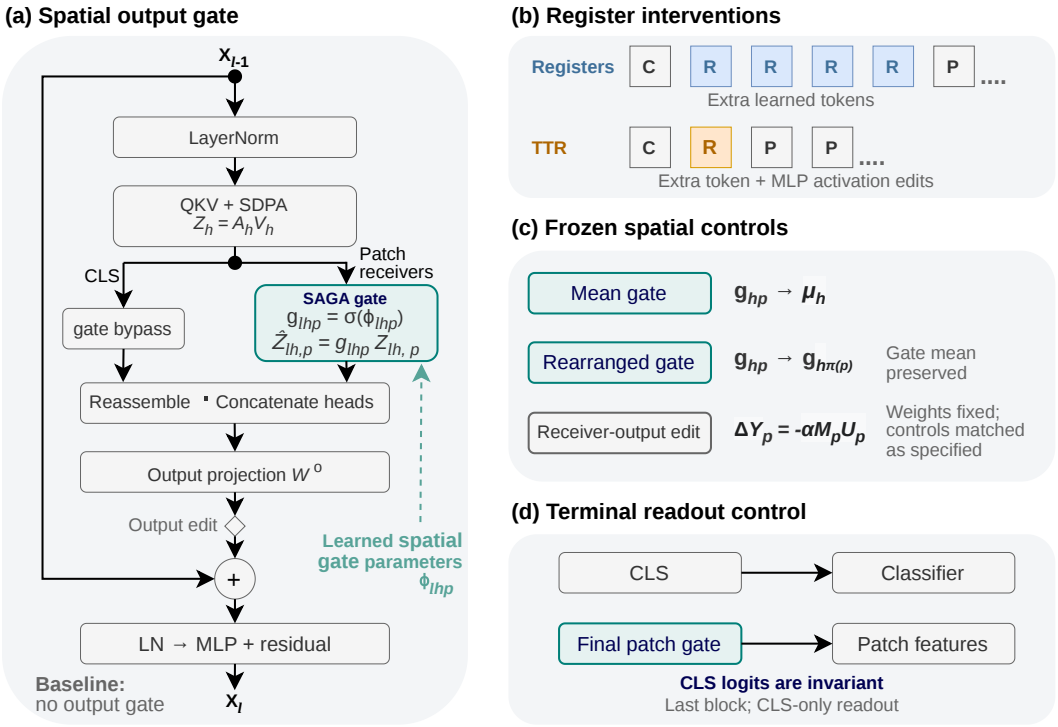

Figure 2: **SAGA and the frozen controls used to separate spatial structure from functional effect.** **(a)** SAGA gates each attention head after scaled dot-product attention and before the output projection: $g_{\ell hp} = \sigma(\phi_{\ell hp})$ scales patch receiver p, while CLS bypasses the gate. The gate is input-independent and indexed by block, head and patch position. **(b)** Trained registers (Darcet et al., 2023) add learned tokens, whereas test-time registers (TTR) (Jiang et al., 2025) redirect identified activation to an added token at inference. **(c)** Frozen controls replace each head's spatial gate by its mean, rearrange its spatial assignment while preserving the mean, or apply matched receiver-output perturbations; model weights remain fixed. **(d)** In the terminal block, changing a patch-only gate can alter patch features but not CLS logits under a CLS-only classifier because no later operation mixes patch rows into CLS.

High-norm tokens and spatial organization. Motivated by prior work identifying unusually high-norm patch tokens as ViT representation artifacts (Darcet et al., 2023), we operationalize a high-norm token as a patch representation whose ℓ_2 norm exceeds a fixed, baseline-calibrated threshold. For image i, let X_{ip} denote the patch representation at position $p \in \{1, \ldots, N\}$ at a declared feature stage, excluding CLS and register tokens. We define

$$n_{ip} = \|X_{ip}\|_2, \qquad b_{ip}(\tau) = \mathbf{1}[n_{ip} > \tau], \qquad c_i = \sum_{p=1}^{N} b_{ip}, \tag{3}$$

where c_i is the number of high-norm tokens in image i. Across M images,

$$f_p = \frac{1}{M} \sum_{i=1}^{M} b_{ip}, \qquad q_p = \frac{f_p}{\sum_{j=1}^{N} f_j}, \tag{4}$$

where f_p measures absolute prevalence at position p, while q_p measures its share of the spatial distribution conditional on a high-norm token occurring. This separation lets us distinguish fewer high-norm tokens from a change in where they occur.

For each architecture, training recipe and feature stage, the primary threshold τ_{cal} is derived from the designated baseline as the median, over calibration images, of the per-image $\text{median}(n_i) +$

$5\ \mathrm{MAD}(n_i)$ threshold, then held fixed across methods within that cell. Because the fixed threshold is sensitive to changes in absolute feature scale, we additionally report per-image median $+5\,\mathrm{MAD}$ exceedances as a relative-scale robustness check. All new patch diagnostics use the output of transformer block 11 ('s11_out'), placing the measurement upstream of the terminal SAGA gate; the historical final-block pre-LayerNorm stage is retained only as a documented control.

Controlled spatial interventions. To test whether spatial structure is functionally meaningful rather than merely visible in the maps, we intervene on trained models while keeping all weights fixed. The controls separate three factors: average gate strength, spatial arrangement and coordinate-specific receiver sensitivity (Fig. 2c). For each SAGA head, we write

$$\mu_{\ell h} = \frac{1}{N}\sum_{p} g_{\ell h p}, \qquad \delta_{\ell h p} = g_{\ell h p} - \mu_{\ell h}, \qquad g^{\pi}_{\ell h p} = \mu_{\ell h} + \delta_{\ell h, \pi(p)}. \tag{5}$$

The *mean* control replaces every spatial gate in a head by $\mu_{\ell h}$, removing spatial variation while preserving its average attenuation. A rearrangement instead preserves the same gate values and mean but changes their coordinate assignment, using one permutation shared across heads. We apply these single-block edits at blocks 8 and 9; within-ring and dihedral rearrangements provide structured controls. Because different rearrangements need not induce equal update magnitudes, responses are stratified by the norm of the resulting attention-update change.

To test coordinates independently of SAGA's learned gates, we perturb the effective patch attention update after any native gate and before residual addition. For mask M,

$$\Delta Y^{(M)}_{ip} = -\alpha_{iM}\mathbf{1}[p \in M]\, U_{ip}, \qquad \alpha_{iM} = \frac{\epsilon \min_{M'} \|M' \odot U_i\|_F}{\|M \odot U_i\|_F}, \tag{6}$$

so perturbation energy is matched across masks within each image and checkpoint. The primary mask contains the 16 highest-prevalence coordinates selected on the disjoint discovery split; ten fixed controls match its boundary-ring composition. We use $\epsilon = 0.10$ for the reported frozen coordinate intervention. Blocks 8 and 9 were fixed before the frozen analysis, although their choice was informed by earlier exploratory correlations.

Readouts and uncertainty. Functional effect depends on the consumer of the representation, so we evaluate classifier and patch-level readouts separately. For intervention v, classifier sensitivity is measured by the paired change in true-label negative log likelihood (NLL),

$$\Delta^{(v)}_i = -\log p_v(y_i \mid x_i) + \log p_0(y_i \mid x_i), \tag{7}$$

where p_0 and p_v denote the native and intervened predictions, respectively and $\Delta^{(v)}_i > 0$ indicates degraded classifier confidence. Patch-level utility is measured independently by exact-transform correspondence between ℓ_2-normalized patch descriptors under deterministic geometric transforms with known patch mappings. The terminal-gate control (Fig. 2d) provides an exact readout separation: modifying the final patch-only gate can change patch features but cannot change CLS logits under a CLS-only classifier, because no later operation mixes patch rows into CLS.

Frozen-edit effects are paired by image and reported with 95% image-level bootstrap confidence intervals. Checkpoint-to-checkpoint variation is reported separately and is not replaced by image-level uncertainty. Spatial-map comparisons use structured spatial permutations rather than iid-coordinate tests; full resampling and correspondence details are provided in the Appendix.

3.2 Cohort and Evaluation Protocol

We study 21 eligible 300-epoch ImageNet-1K (Deng et al., 2009) conditions spanning ViT-S/16 and ViT-B/16: eight Baselines, eight SAGA models and five trained-register models. Eligibility requires completion of the declared final checkpoint and passing audited architecture, recipe and checkpoint-provenance checks; incomplete or invalid runs are excluded. These yield three matched Baseline–SAGA architecture–recipe cells, defined by the audited cohort rather than outcome selection. Training follows a DeiT-style recipe (Touvron et al., 2021a), with mixing or true no-mixing; mixing uses MixUp (Zhang et al., 2017) and CutMix (Yun et al., 2019). Twelve conditions are seed-controlled reruns and nine are audited legacy runs. Test-time registers are inference edits of

Table 1: **SAGA reduces high-norm-token prevalence while improving observed ImageNet-1K accuracy.** B/S denotes Baseline/SAGA, n paired Baseline–SAGA checkpoints and Reg. Δ the trained-register top-1 difference from its matched baseline. SAGA Δ is the paired mean$\pm$SE across eligible checkpoints; high-norm-token counts use the fixed baseline-calibrated threshold.

Cell	n	Top-1 B/S	SAGA Δ	Reg. Δ	High-norm/img B $\rightarrow$ S
ViT-S / Mix	4	78.720 / 79.176	$+0.456 \pm 0.212$	-0.381 ($n = 2$)	16.480 $\rightarrow$ 8.401
ViT-S / NoMix	2	73.035 / 73.362	$+0.327 \pm 0.007$	-0.349 ($n = 2$)	4.051 $\rightarrow$ 1.433
ViT-B / Mix	2	76.925 / 77.363	$+0.438 \pm 0.286$	-0.610 ($n = 1$)	8.404 $\rightarrow$ 4.534

Table 2: **Downstream evaluation of SAGA.** Aircraft results are means over three paired fine-tuning seeds (Δ shown as mean$\pm$SE); COCO and ADE20K are single-run evaluations. Full downstream metrics and all evaluated cells are reported in the Appendix.

Dataset	Backbone	Metric	Baseline	SAGA	Δ
FGVC-Aircraft	ViT-S/16	Top-1	73.07	74.03	$+0.96 \pm 0.44$
FGVC-Aircraft	ViT-B/16	Top-1	72.53	73.62	$+1.09 \pm 0.42$
COCO	ViT-B/16	AP	34.91	35.34	$+0.43$
ADE20K	ViT-B/16	mIoU_{ms}	43.97	43.97	≈ 0.00

existing Baselines, not trained checkpoints (Fig. 2b). We separate trained-system comparisons from frozen mechanistic tests and use disjoint discovery, calibration and evaluation splits for hypothesis formation, calibration and final reporting. Appendix A reports cohort details and exclusions.

4 Results

We first evaluate SAGA as a trained system, then use frozen analyses to determine what its reduction in high-norm tokens does and does not imply. We characterize their attention and spatial structure, test coordinate-specific classifier effects and compare classifier and patch-level readouts.

4.1 SAGA Reduces High-Norm Tokens While Improving Accuracy

Across all three audited 300-epoch ImageNet-1K architecture–recipe cells, SAGA yields fewer fixed-threshold high-norm tokens and higher observed mean top-1 accuracy than the corresponding ViT Baseline (Fig. 1c; Table 1). The paired changes are $+0.456$ pp with 49.0% fewer high-norm tokens for ViT-S/Mix, $+0.327$ pp with 64.6% fewer for ViT-S/NoMix, and $+0.438$ pp with 46.0% fewer for ViT-B/Mix. The four ViT-S/Mix top-1 differences are -0.126, $+0.570$, $+0.490$, and $+0.890$ pp.

Because fixed-threshold counts are scale-sensitive, we also evaluate per-image median $+5\,\text{MAD}$ exceedances. They change from 16.276 to 14.435 events/image for ViT-S/Mix and 3.094 to 2.822 for ViT-S/NoMix, but increase from 9.921 to 16.929 for ViT-B/Mix. Thus, the fixed-threshold result reflects absolute-scale event reduction rather than a universal reduction of within-image relative tails.

Stronger suppression does not itself determine classification change. In the matched trained-register comparisons, fixed-threshold counts decrease more strongly while top-1 differences are negative in all three cells (Table 1); with only $n = 1$–2 register checkpoints per cell and legacy provenance, we treat this descriptively rather than as a method ranking. Likewise, two ViT-S/Mix TTR operating points remove roughly 90% of fixed-threshold high-norm tokens for about a 0.2 pp top-1 decrease.

Beyond ImageNet, SAGA improves FGVC-Aircraft top-1 on both evaluated backbones and COCO overall AP, while multi-scale ADE20K mIoU is essentially unchanged (Table 2). Aircraft uses three paired fine-tuning seeds, whereas COCO and ADE20K are single-run evaluations and therefore provide supporting rather than replicated cross-task evidence.

These results motivate the mechanistic question that follows: what distinguishes high-norm tokens and does their spatial organization carry functional information?

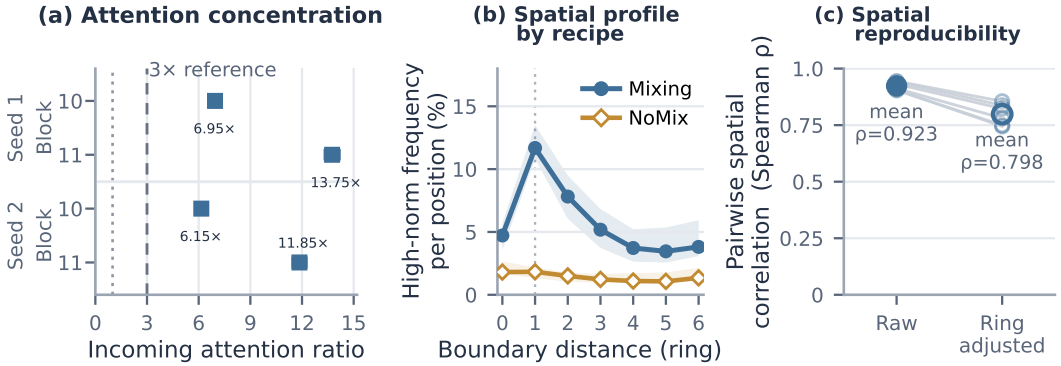

Figure 3: **High-norm tokens are reproducible attention sinks. (a)** High-norm positions receive 6.15–13.75× more incoming attention per token than non-exceedance positions in two fresh ViT-S/16 mixing checkpoints at two late blocks; the dashed line marks the 3× reference. **(b)** High-norm frequency varies systematically with boundary distance: all 24 run–stage profiles from six Mixing Baselines peak at ring 1, whereas none of eight profiles from two true-NoMix Baselines do. **(c)** Pairwise mixing prevalence-map correlations remain high after removing the mean boundary-ring profile, indicating reproducible spatial structure beyond the coarse ring effect.

4.2 High-Norm Tokens Are Attention Sinks with Reproducible Spatial Structure

Because stronger high-norm-token suppression does not necessarily yield higher accuracy, we next ask what these tokens represent mechanistically. In both fresh ViT-S/16 mixing baselines, high-norm patch keys receive 6.15–13.75× more incoming attention than other patch keys, averaging the original softmax attention over heads and patch queries across the two measured late blocks (Fig. 3a). All four 95% image-bootstrap intervals remain above the 3× reference, with the smallest lower bound at 6.02×. Thus, within this cohort, high-norm tokens are strong *attention sinks*.

Their aggregate locations are also highly structured. Across six eligible Mixing Baselines, all 24 run–stage profiles peak at boundary ring 1, one patch inside the image boundary, whereas none of the eight profiles from two true-NoMix Baselines do (Fig. 3b). At 's11_out', the fresh ViT-S/16 mixing prevalence maps are highly consistent across checkpoints (mean pairwise Spearman $\rho = 0.923$) and substantial agreement remains after removing the mean boundary-ring profile ($\rho = 0.798$; Fig. 3c). Yet position explains only 2.21–2.33% of image-level high-norm-token variance, showing that a stable population-level spatial signature can coexist with substantial image-specific variation.

Reproducibility, however, does not by itself identify a learned coordinate-specific mechanism. The scale-matched control in Fig. 1b globally rescales baseline patch norms to match SAGA's high-norm-token rate and nevertheless reproduces much of the preserved spatial profile. The recipe-linked pattern also extends beyond our trained cohort, holding in 13 of 16 applicable public model-stage evaluations, with three counterexamples. Together, these results establish a reproducible spatial phenomenon while showing that neither its persistence nor its recipe association is sufficient evidence that particular coordinates are functionally privileged.

4.3 Spatial Reproducibility Does Not Imply Classifier Sensitivity

The reproducible structure in Fig. 3 motivates a stronger hypothesis: the classifier may depend on where high-norm activity concentrates. We test this with frozen interventions separating average gate strength from spatial assignment. Replacing each SAGA head's gate by its mean (C1), or rearranging the same values across coordinates (C2), yields no reproducible classifier response across two fresh ViT-S/Mix checkpoints at Blocks 8–9 (Fig. 4a). Point estimates are at most 3.5×10^{-4} nats/image and remain non-replicating after stratification by induced attention-update norm.

We next attenuate the receiver-side update at the 16 highest-prevalence coordinates and compare them with ten boundary-ring-matched controls under per-image energy matching (Fig. 4b). The

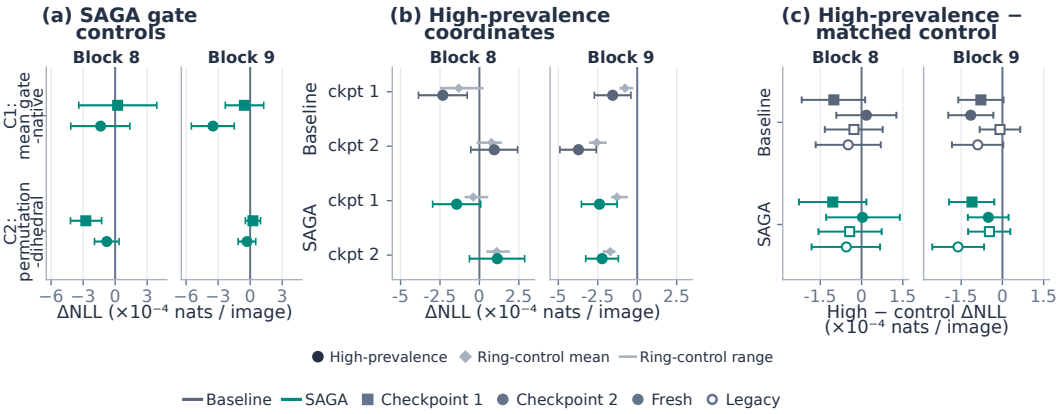

Figure 4: **Spatial reproducibility does not imply coordinate-specific classifier sensitivity. (a)** Mean-gate (C1) and gate-arrangement (C2) controls at blocks 8–9. **(b)** High-prevalence coordinates versus boundary-ring-matched controls under per-image energy matching. **(c)** The high-prevalence-minus-control contrast (C3) does not replicate across fresh Baseline or SAGA checkpoint pairs. Error bars show 95% image-bootstrap CIs; zero denotes no NLL difference.

resulting high-prevalence-minus-control contrast (C3; Fig. 4c) again fails to replicate across fresh Baseline or SAGA checkpoint pairs. At Block 9, one checkpoint per method individually excludes zero, but neither effect replicates.

Thus, reproducible spatial structure does not imply reproducible coordinate-specific CLS sensitivity under the tested Block 8–9 interventions. We retain these pre-specified blocks rather than retargeting them after observing the Block 10–11 attention-concentration results, and next examine a patch-level consumer.

4.4 Functional Relevance Depends on the Readout

The terminal SAGA gate provides an exact architectural readout control. Changing this patch-only gate alters the final-block patch diagnostic (Fig. 5a), yet leaves CLS logits exactly unchanged across all eight SAGA checkpoints (Fig. 5b): the maximum absolute logit difference is 0, with top-1 agreement = 1.000. After the terminal patch-only gate, no operation mixes patch rows back into the CLS representation used by the classifier.

We next evaluate a consumer that directly uses patch features. In the two fresh ViT-S/Mix Baselines at 's11_out', high-norm positions have 47–55% exact-transform correspondence versus 93–95% for other positions across horizontal flip and one- and two-patch shifts (Fig. 5c). Across the six checkpoint–transform rows, the mean difference is -43.8 pp, with every 95% image-bootstrap CI excluding zero; images with more high-norm positions also tend to have poorer correspondence (mean Spearman $\rho = -0.440$).

Together, these results show that functional relevance is *readout-dependent*: the same diagnostic can be exactly invisible to a CLS-only classifier while being associated with poorer patch-level correspondence. Because high-norm positions are themselves boundary-biased, this unadjusted association may partly reflect spatial location and is interpreted descriptively rather than causally.

5 Limitations

Our conclusions are bounded by the evaluated checkpoints and interventions. Several ImageNet cells contain only two paired checkpoints, and trained-register comparisons include legacy runs with limited replication; the results therefore support within-cell rather than universal architecture claims. FGVC-Aircraft uses three paired fine-tuning seeds, whereas COCO and ADE20K are single-run evaluations and provide supporting rather than replicated cross-task evidence.

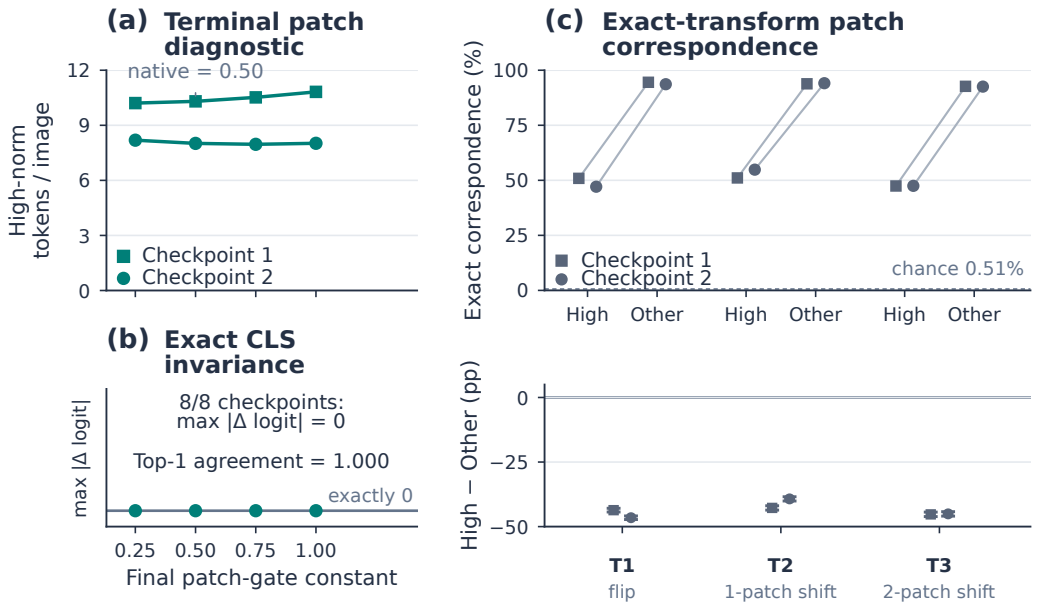

Figure 5: **Functional relevance of the high-norm diagnostic depends on the readout.** **(a)** Changing the terminal patch-gate constant alters the final-block patch diagnostic. **(b)** The same edits leave CLS logits exactly invariant across all eight SAGA checkpoints: max $|\Delta\text{logit}| = 0$ and top-1 agreement $= 1.000$. **(c)** In two fresh ViT-S/Mix Baselines at 's11_out', high-norm positions are associated with substantially lower exact-transform correspondence than other positions under flip and one- and two-patch shifts. Error bars show 95% image-bootstrap CIs.

The frozen tests cover selected late blocks, checkpoints and perturbations. Their non-replicating contrasts do not establish equivalence or generalize to other layers, interventions or readouts. Correspondence is evaluated only on two fresh ViT-S/Mix Baselines, and its deficit may partly reflect the boundary bias of high-norm positions; we therefore interpret it descriptively rather than causally.

The public-model audit also contains counterexamples and heterogeneous training metadata, supporting partial rather than universal external scope. Broader replication across architectures, training regimes and downstream consumers remains necessary.

6 Conclusion

We introduced SAGA, a lightweight spatially indexed attention gate that adds no tokens and only a negligible number of parameters. Across the audited ImageNet architecture–recipe cells, SAGA improves observed mean classification accuracy while reducing fixed-threshold high-norm-token prevalence, with supporting evidence from additional downstream evaluations.

The mechanistic analyses show that prevalence, spatial organization and downstream functional effect should not be treated as interchangeable. High-norm tokens are strong attention sinks with reproducible spatial structure, yet global norm rescaling reproduces much of this structure and matched frozen interventions do not reveal reproducible coordinate-specific CLS sensitivity in the tested blocks. Conversely, the same patch-level diagnostic can be exactly invisible to a CLS-only classifier while remaining associated with poorer local correspondence.

Together, these results motivate evaluating representation diagnostics against the computation that actually consumes them. Future work should test whether these distinctions persist across larger backbones, broader training regimes and patch-dependent downstream readouts.

AI USE STATEMENT

Generative AI tools were used in this work as research and software-development assistants. They were used to provide feedback on experimental design, suggest candidate controls and robustness checks, discuss alternative explanations of observed results and assist with the interpretation and presentation of experimental findings. AI-assisted tools were also used to help implement, edit, debug and review portions of the experimental, analysis and plotting code. The research questions, final experimental protocol, inclusion and exclusion decisions, choice of reported analyses and scientific conclusions were determined by the authors after reviewing the available empirical evidence.

Generative AI tools were additionally used for literature discovery and synthesis, manuscript organization, drafting and language editing and the presentation of figures and tables. The authors independently checked cited sources, verified reported numerical and statistical results against the underlying experiment outputs and reviewed AI-assisted code before using its outputs as evidence. Experiments and analyses reported in the paper were executed and inspected by the authors. The authors reviewed all AI-assisted material and take responsibility for the final text, claims, code, figures and other artifacts in this work.

ETHICS STATEMENT

This work studies the internal representations and attention behavior of vision transformers and does not involve human-subject experiments or the collection of new personal or sensitive data. The empirical study uses established computer-vision research benchmarks and publicly released pretrained models under their respective research-use conditions; no new person-level annotations or datasets are introduced.

The proposed analyses are intended to improve the interpretation of representation diagnostics in vision models. In particular, we avoid treating high-norm tokens, spatial patterns or their suppression as evidence of representation quality or causal importance without corresponding functional tests. The reported correspondence results are likewise not intended as measures of semantic quality, fairness or model safety. More broadly, the models and datasets considered here may inherit limitations or biases from their training data, which this work does not attempt to measure or remove. The authors take responsibility for the appropriate presentation and interpretation of these results.

REPRODUCIBILITY STATEMENT

We provide an anonymous supplementary code archive containing the SAGA implementation, experiment configurations, evaluation and analysis scripts and plotting utilities used to produce the reported results and figures. The Method section specifies the model intervention, diagnostic definitions, feature stages, frozen spatial controls and statistical procedures used in the main experiments. The appendix provides additional implementation details, checkpoint and experiment provenance, data-split and threshold definitions, complete intervention results, robustness analyses, external-model results and the full secondary and downstream evaluations.

Where applicable, experiment configurations, seeds, and analysis settings are recorded in the supplied configuration and result files; legacy checkpoints whose seed provenance is unavailable are explicitly identified rather than treated as matched seeded runs. The supplementary material also contains the scripts needed to regenerate the reported tables and figures from the corresponding experimental outputs. Upon publication, we will release the code and reproducibility materials through a public GitHub repository to facilitate independent replication and extension of the work.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep VIT features as dense visual descriptors, 12 2021. URL https://arxiv.org/abs/2112.05814.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sbusw6LD41.

Yinjie Chen, Zipeng Yan, Chong Zhou, Bo Dai, and Andrew F. Luo. Vision Transformers with Self-Distilled Registers, 5 2025. URL https://arxiv.org/abs/2505.21501.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 9 2023. URL https://arxiv.org/abs/2309.16588.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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, 10 2020. URL https://arxiv.org/abs/2010.11929.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers, 6 2025. URL https://arxiv.org/abs/2506.08010.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*, September 2014. URL https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-Grained visual Classification of aircraft, 6 2013. URL https://arxiv.org/abs/1306.5151.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features Without Supervision. *Transactions on Machine Learning Research*, 2024. URL https://mlanthology.org/tmlr/2024/oquab2024tmlr-dinov2/.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: non-linearity, sparsity, and Attention-Sink-Free, 5 2025. URL https://arxiv.org/abs/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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26328–26337, June 2026.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers, 6 2021. URL https://arxiv.org/abs/2106.10270.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on*

Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/touvron21a.html.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with Image Transformers, 3 2021b. URL https://arxiv.org/abs/2103.17239.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DEIT III: Revenge of the VIT. *Lecture notes in computer science*, pp. 516–533, 1 2022. doi: 10.1007/978-3-031-20053-3_30. URL https://doi.org/10.1007/978-3-031-20053-3_30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 6 2017. URL https://arxiv.org/abs/1706.03762.

Haoqi Wang, Tong Zhang, and Mathieu Salzmann. SINDER: Repairing the singular defects of DINOV2. *Lecture notes in computer science*, pp. 20–35, 9 2024. doi: 10.1007/978-3-031-72667-5_2. URL https://doi.org/10.1007/978-3-031-72667-5_2.

Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers, 1 2024. URL https://arxiv.org/abs/2401.02957.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features, 5 2019. URL https://arxiv.org/abs/1905.04899.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization, 10 2017. URL https://arxiv.org/abs/1710.09412.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Table 3: **Eligible 300-epoch ImageNet-1K cohort.** Counts refer to independently trained conditions; inference-time edits such as TTR are not included.

Cell	Baseline	SAGA	Registers	Total
ViT-S / Mix	4	4	2	10
ViT-S / NoMix	2	2	2	6
ViT-B / Mix	2	2	1	5
Total	8	8	5	21

A Experimental Setup, Cohort, and Measurement Definitions

This appendix summarizes the experimental cohort, data partitions, feature-stage conventions and intervention definitions used throughout the supplementary analyses. Exhaustive checkpoint-level provenance and machine-readable analysis records are provided with the supplementary artifacts.

A.1 Cohort and Provenance

The final 300-epoch ImageNet-1K cohort contains 21 eligible independently trained conditions: eight Baselines, eight SAGA models, and five trained-register models. Twelve conditions are seed-controlled reruns and nine are audited eligible legacy runs. Table 3 summarizes the architecture–recipe composition.

The legacy and seed-controlled conditions are retained as distinct provenance classes rather than treated as interchangeable training replicates. The two additional ViT-S/NoMix trained-register checkpoints were completed after the original frozen-analysis manifest was constructed. They extend the final trained-system comparison but do not alter any frozen split, threshold, mask, permutation, intervention layer, or analysis decision used in the inference studies.

Eligibility is determined from the verified training configuration and completed checkpoint rather than from historical directory names. Incomplete or invalid training runs contribute to no reported aggregate, and multiple saved epochs or evaluation variants from the same training run are not counted as additional conditions. Test-Time Registers (TTR) are treated separately throughout because they are inference-time edits of an existing Baseline checkpoint rather than independently trained models.

A.2 Data partitions, feature stages, and statistical evaluation

We use disjoint discovery, calibration, and evaluation partitions containing 10,000, 2,000, and 10,000 ImageNet-1K validation images, respectively. Discovery is used for hypothesis-dependent choices such as the high-prevalence coordinate mask, calibration for fixed-threshold estimation, and evaluation for final frozen analyses. Expensive analyses use fixed subsets of the evaluation partition: 2,000 images for correspondence and 1,000 for attention extraction.

For each architecture–recipe cell and feature stage, the fixed high-norm threshold is calibrated from the designated Baseline as the median, over calibration images, of the per-image $\mathrm{median}(n_i) + 5\ \mathrm{MAD}(n_i)$ statistic and then held fixed across methods within that comparison. We additionally report per-image $\mathrm{median} + 5\ \mathrm{MAD}$ exceedances as a relative-scale robustness check. Unless stated otherwise, new patch analyses use the output of transformer Block 11 ('s11_out'); the historical final-block pre-LayerNorm stage ('s12_pre_norm') is retained only as a documented control.

For the frozen coordinate intervention, the 16 highest-prevalence positions are selected on the discovery partition and compared with ten fixed controls having the same boundary-ring composition. Intervention energy is matched within each image and checkpoint. Frozen effects are paired by image and reported with 95% image-level bootstrap confidence intervals from 10,000 resamples (seed 0), with images as the resampling unit.

For exact-transform correspondence, ℓ_2-normalized patch descriptors are compared under horizontal flip and one- and two-patch translations with known coordinate mappings. On the 14×14 patch

grid, random exact correspondence is $1/196$. Spatial-map comparisons use structured spatial permutations rather than independent coordinate-wise tests.

B Spatial Structure and Scale Controls

The main text reports the principal spatial findings at 's11_out'. Here we provide the complete checkpoint-level maps and the controls used to separate spatial organization from overall high-norm-token prevalence. The analyses in this section are descriptive: they establish reproducible spatial structure and its dependence on the training condition, but do not by themselves establish coordinate-specific functional importance.

B.1 Complete Spatial Structure

Figure 6 expands the spatial analysis to the four eligible ViT-S/Mix Baseline–SAGA pairs. Across the four Baseline checkpoints, the mean number of high-norm tokens per image is 20.37, 17.78, 17.72 and 16.54; the corresponding SAGA checkpoints yield 11.46, 12.59, 7.53 and 9.48. Despite this reduction in absolute prevalence, the remaining events retain a highly reproducible spatial organization.

These counts differ from the pooled values in Table 1 because that table retains the historical 's12_pre_norm' trained-system diagnostic and canonical threshold, whereas Fig. 6 uses the disjoint evaluation split, the primary 's11_out' stage and its stage-specific calibrated threshold.

Across the six Baseline checkpoint pairs, the mean spatial Spearman correlation is $\rho = 0.923$ and remains $\rho = 0.798$ after adjustment for boundary-ring structure. The recipe comparison provides a complementary view: all 24 Mixing run–stage profiles peak one patch inside the image boundary, whereas none of the eight NoMix profiles do. Thus, the observed pattern is not explained solely by one unusually structured checkpoint. At the same time, position accounts for only a modest fraction of image-level event variability, so reproducibility of the population map should not be interpreted as spatial position determining individual high-norm events.

B.2 Scale and Recipe Controls

A reduction in threshold exceedances can itself change the apparent spatial map. We therefore compare the observed profiles with a scale-only control in Figure 7. For each Baseline condition, the recorded patch norms are globally rescaled to approximately match the corresponding SAGA high-norm-token rate; the same scalar is then applied without altering spatial ordering.

The scale-only control reproduces a substantial portion of the observed Baseline–SAGA spatial similarity. This limits what can be inferred from a lower high-norm-token count together with a preserved spatial map: much of that combination can arise from changing absolute norm scale while retaining the underlying spatial ordering. The recipe comparison nevertheless remains distinct, with the ring-1 peak consistently present in the Mixing conditions and absent from the true-NoMix conditions. We therefore treat overall prevalence, spatial organization, and functional sensitivity as separate quantities; the latter is tested directly in Appendix C.

C Frozen Spatial Interventions

The reproducible spatial patterns in Appendix B motivate a stronger question: does the particular spatial arrangement carry distinct functional importance for the frozen classifier? We test this in two complementary ways. First, we modify the learned SAGA gate while keeping the trained checkpoint fixed. Second, we perturb high-prevalence patch coordinates and compare their classifier response with boundary-ring-matched controls. All interventions are applied at the originally selected Blocks 8 and 9; the intervention sites are not retargeted after observing later-layer attention-sink measurements.

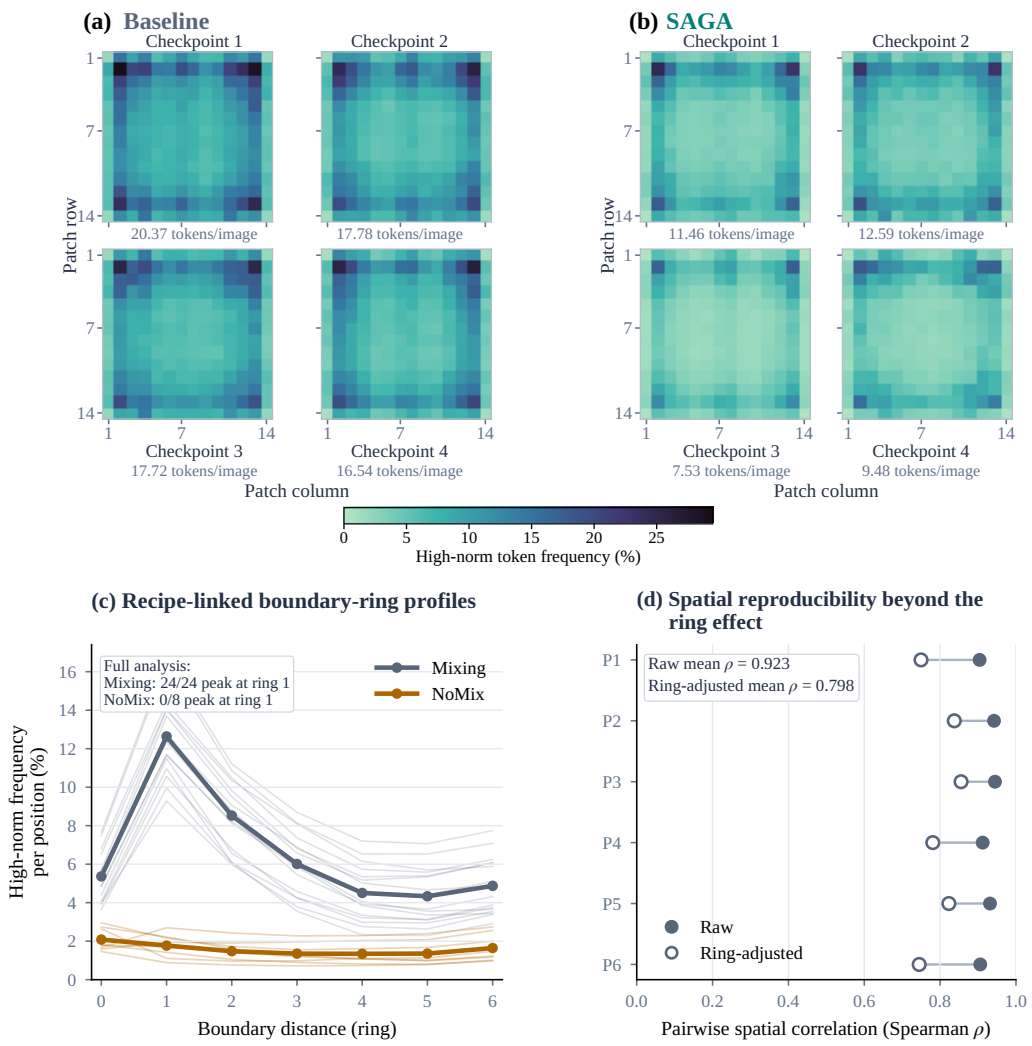

Figure 6: **Complete spatial evidence.** (a and b) High-norm-token prevalence maps for four paired ViT-S/16 checkpoints trained with mixing, evaluated at 's11_out'. Absolute prevalence decreases under SAGA, while substantial spatial organization remains. (c) Boundary-ring profiles across the evaluated stages and checkpoints. All 24 Mixing run–stage profiles peak at ring 1, whereas none of the eight true-NoMix profiles do. (d) Pairwise spatial reproducibility for the four fresh Mixing Baselines. Mean Spearman correlation is $\rho = 0.923$ before adjustment and $\rho = 0.798$ after removing the boundary-ring component.

C.1 Learned Gate Arrangement

Figure 8 gives the complete frozen gate-arrangement results for the two fresh ViT-S/Mix SAGA checkpoints. For each edited block, the native gate is compared with its per-head spatial mean and with frozen spatial rearrangements. We summarize the two primary contrasts as

$$C1 = \mathrm{NLL}(\text{mean gate}) - \mathrm{NLL}(\text{native}), \tag{8}$$

$$C2 = \mathrm{NLL}(\text{permutation}) - \mathrm{NLL}(\text{dihedral}), \tag{9}$$

where the dihedral term denotes the nonidentity structured control family.

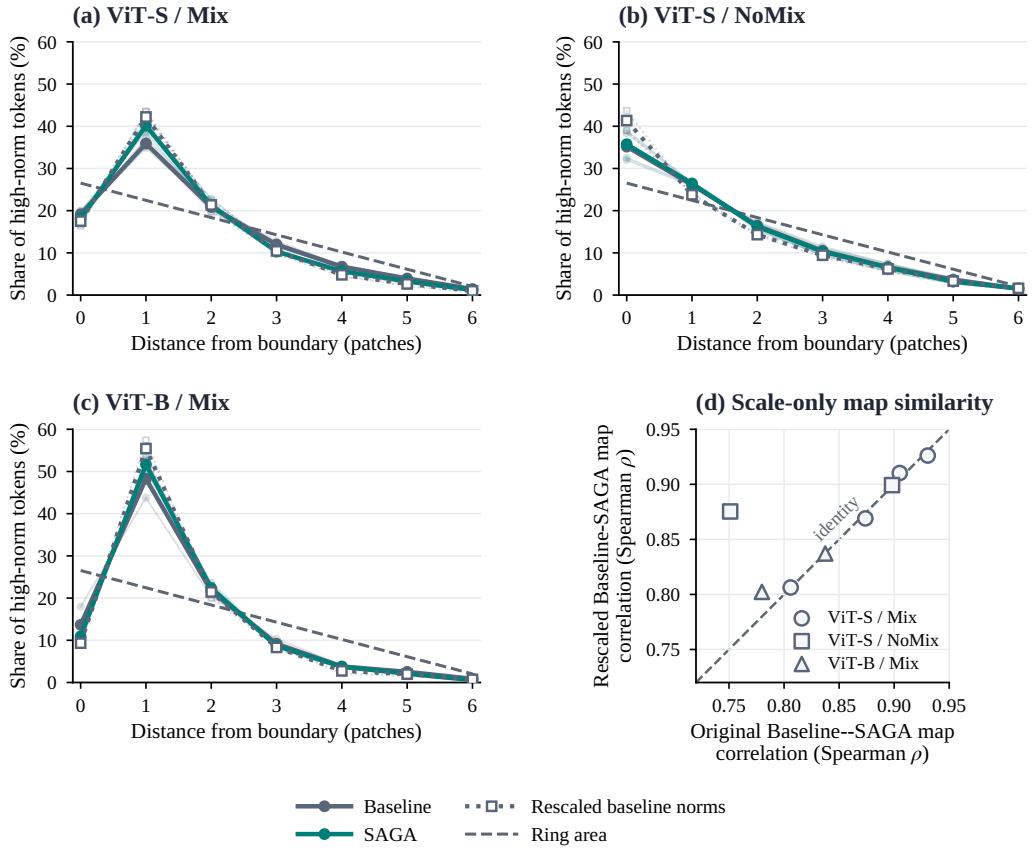

Figure 7: **Scale and recipe controls for spatial prevalence.** (a–c) Boundary-ring profiles for ViT-S/Mix, ViT-S/NoMix, and ViT-B/Mix, comparing Baseline, SAGA, rescaled Baseline norms, and the corresponding ring-area reference. (d) Original Baseline–SAGA spatial correlation versus the correlation obtained after globally rescaling the recorded Baseline norms to match SAGA prevalence. The rescaling is a statistical control on stored norm values, not a modified forward model or classifier.

The individual effects are small, with the primary point estimates bounded at approximately 3.5×10^{-4} nats/image in magnitude, and neither $C1$ nor $C2$ satisfies the cross-checkpoint replication criterion. Stratifying the rearrangement response by induced update magnitude does not reveal a reproducible effect that is hidden in the aggregate comparison. These results therefore do not support the hypothesis that preserving the fine learned spatial arrangement is required for classifier performance in the tested checkpoints and blocks. They do not establish equivalence, nor do they imply that spatial degrees of freedom were irrelevant during training.

C.2 High-Prevalence Coordinates

We next ask a related but distinct question: whether coordinates with high population prevalence respond differently to a controlled attenuation than geometrically matched coordinates. The high-prevalence mask contains 16 patch positions selected from the disjoint discovery split. It is compared with ten fixed control masks having exactly the same boundary-ring composition. For every image and candidate mask, attenuation strength is normalized so that the injected attention-update energy is matched within the checkpoint. The primary perturbation strength is $\epsilon = 0.10$; $\epsilon = 0.25$ is reported only as a secondary sensitivity analysis.

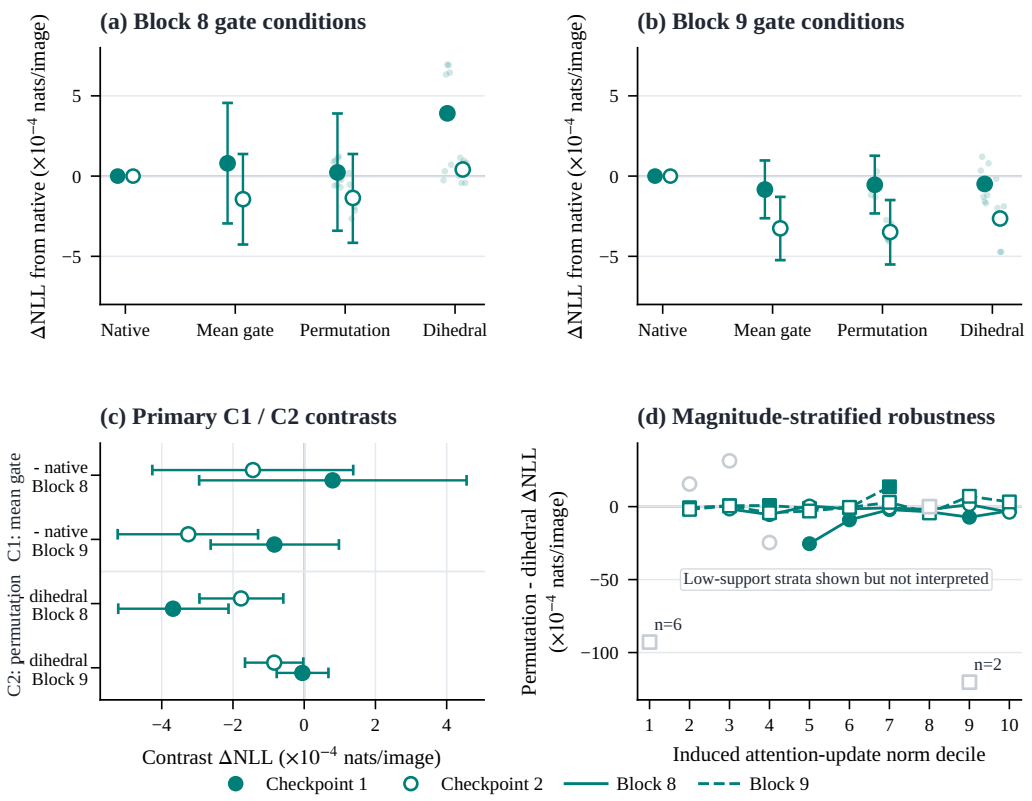

Figure 8: **Complete frozen gate-arrangement interventions.** (a–b) Change in classifier NLL under the completed gate conditions at Blocks 8 and 9 for the two fresh ViT-S/Mix SAGA checkpoints. (c) Primary checkpoint-level contrasts: $C1 = \text{mean gate} - \text{native}$ and $C2 = \text{permutation} - \text{dihedral}$. Neither contrast reproduces consistently across the fresh checkpoint pair. (d) C2 responses stratified by the magnitude of the induced attention update. Low-support strata are retained for completeness but are visually de-emphasized and not interpreted. Error bars denote the corresponding image-level uncertainty where defined; repeated intervention conditions are not independent training replicates.

For method and checkpoint m, the primary comparison is

$$C3_m = \mathbb{E}_i \left[\Delta_i^{\text{high},m} - \frac{1}{10} \sum_{r=1}^{10} \Delta_i^{\text{control}(r),m} \right]. \tag{10}$$

Positive values indicate that attenuation at the high-prevalence coordinates increases NLL more than the mean matched control; negative values indicate a smaller degradation or a larger improvement.

The controls eliminate two immediate alternatives: the compared masks have the same boundary-ring composition, and the injected update magnitude is matched within each image and checkpoint. Nevertheless, $C3$ does not reproduce across the fresh Baseline pair or across the fresh SAGA pair. At Block 9, individual checkpoint intervals can exclude zero, but the effect does not replicate across the corresponding independently trained checkpoints. We therefore interpret the result at the replication level rather than selecting the strongest individual checkpoint.

Together, Figures 8 and 9 separate reproducible spatial prevalence from reproducible classifier sensitivity. Under the declared frozen interventions, neither the fine SAGA gate arrangement nor the high-prevalence coordinate set exhibits a reproducible distinct classifier-level effect. This conclusion is specific to the tested checkpoints, Blocks 8–9, and intervention definitions; the null contrasts are not equivalence tests.

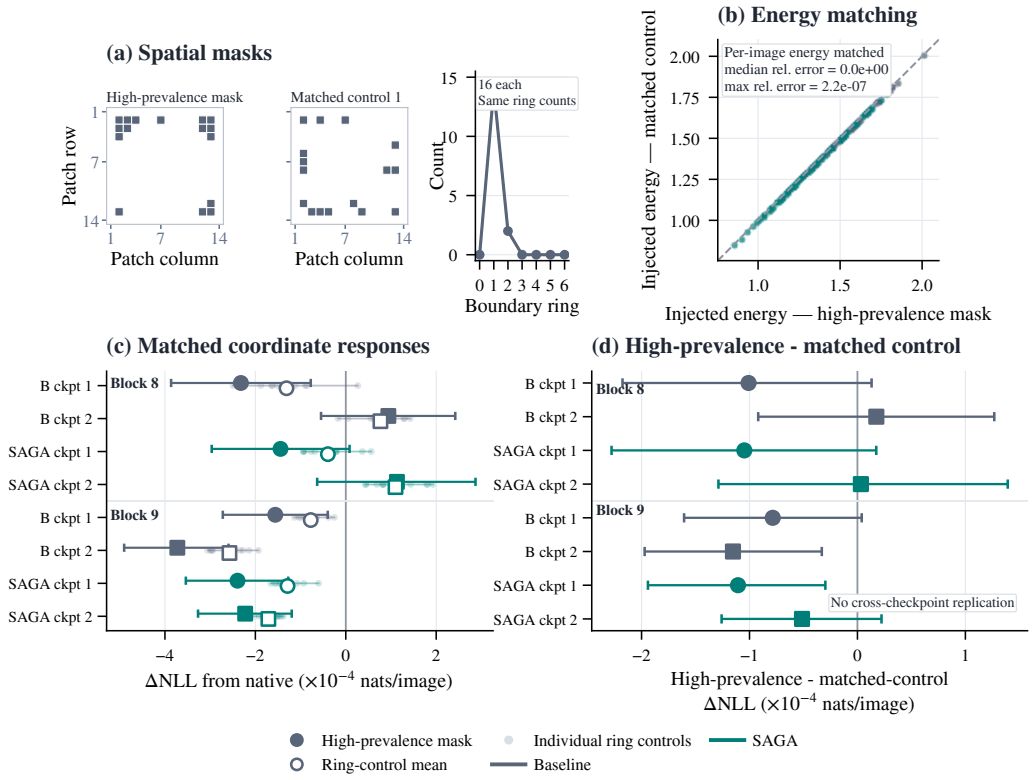

Figure 9: **Complete matched coordinate interventions.** (a) The fixed 16-position high-prevalence mask and boundary-ring-matched control construction. (b) Verification of per-image injected-energy matching. (c) Classifier responses for the high-prevalence mask and the ten fixed ring-matched controls at Blocks 8 and 9. (d) Primary contrast $C3 = \text{high-prevalence} - \text{mean ring control}$ at $\epsilon = 0.10$ (primary perturbation strength). The primary contrast does not replicate across the fresh Baseline pair or across the fresh SAGA pair. Individual checkpoint intervals that exclude zero are retained but are not interpreted as replicated method-level effects.

D Readout Dependence and Patch Correspondence

The frozen spatial interventions in Appendix C do not show a reproducible coordinate-specific classifier effect. This does not imply that the high-norm diagnostic is functionally uninformative for every consumer. We therefore separate two questions: whether patch representations can change while a CLS-only classifier remains exactly fixed, and whether high-norm positions differ under a patch-level geometric readout.

D.1 Terminal Patch-Gate Readout Control

We first vary only the terminal SAGA patch gate while leaving the remainder of each trained checkpoint unchanged. The sweep replaces the terminal gate by constants 0.25, 0.50, 0.75, and 1.00 across all eight SAGA classification checkpoints: four ViT-S/Mix, two ViT-S/NoMix, and two ViT-B/Mix. The frozen verification identifies the 0.50 condition with the native terminal gate for these checkpoints, while 1.00 bypasses its attenuation.

The separation in Figure 10 is architectural, not a failure to detect a small classifier effect. The terminal gate acts only on patch receiver rows after the final attention mixing; the subsequent operations are tokenwise, and the classifier consumes only the CLS representation. Consequently, changing the terminal patch gate cannot propagate the edited patch rows back into the classifier's

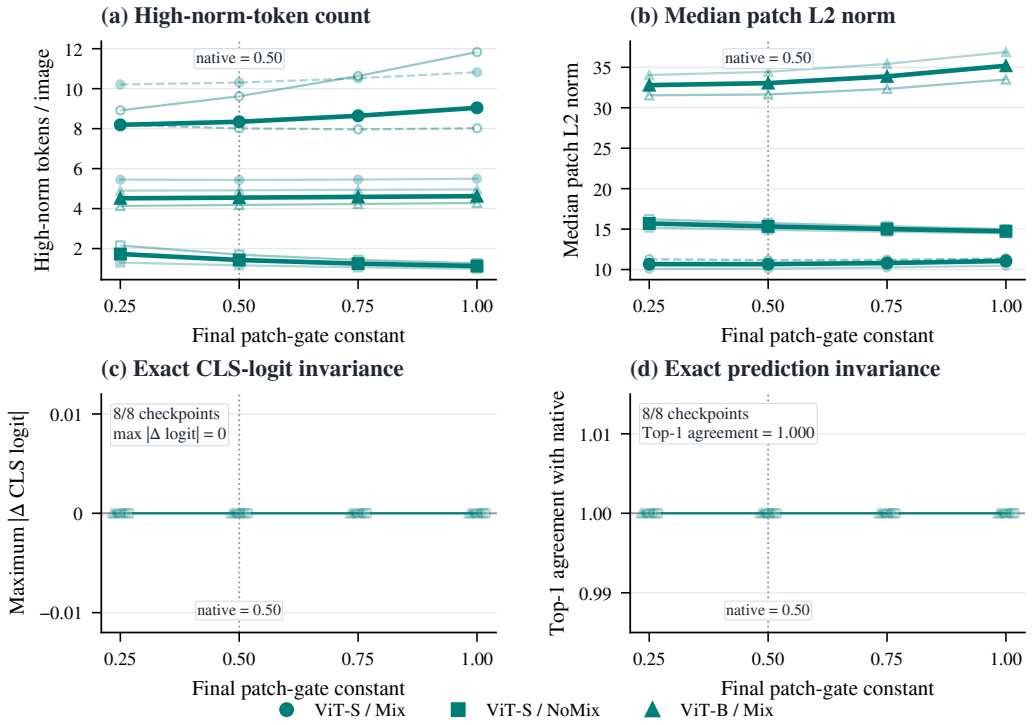

Figure 10: **Complete terminal patch-gate readout control across all eight SAGA checkpoints.** (a–b) Replacing only the terminal patch gate by constants 0.25, 0.50, 0.75, and 1.00 changes the post-edit final-block patch representation, as reflected by the high-norm-token count and median patch L2 norm. The native terminal condition is 0.50, and the high-norm threshold is held fixed across edits within each comparison. (c) Despite these patch-level changes, the maximum absolute change in any CLS classifier logit is exactly zero for every checkpoint and gate constant. (d) Top-1 predictions are correspondingly identical to the native classifier, with agreement 1.000 in every tested condition. This exact invariance applies to the CLS-only classification readout and does not imply invariance of patch-level or dense readouts.

CLS representation. Empirically, this yields

$$\max|\Delta\text{logit}| = 0, \qquad \text{Top-1 agreement} = 1.000$$

for all eight checkpoints and all four gate constants.

Importantly, the patch diagnostics in Figure 10(a–b) are measured downstream of the edited terminal gate. They should not be confused with the paper's primary 's11_out' diagnostic, which lies upstream of this edit and is therefore unchanged by the terminal gate sweep. The control demonstrates that a patch diagnostic can move while the corresponding CLS-only classifier output remains exactly fixed.

D.2 High-Norm Tokens and Exact Patch Correspondence

We next evaluate a consumer that directly uses patch representations. For the two fresh ViT-S/Mix Baseline checkpoints, patch descriptors are taken at 's11_out' and matched under three known geometric transforms: T1 is a horizontal flip, T2 is a one-patch translation, and T3 is a two-patch translation. Because the corresponding patch coordinate is known exactly, random matching on the 14×14 grid has probability $1/196 \approx 0.51\%$.

The diagnostic-group separation is large and consistent across both fresh checkpoints and all three transforms: high-norm positions retain only 47–55% exact correspondence, whereas other positions

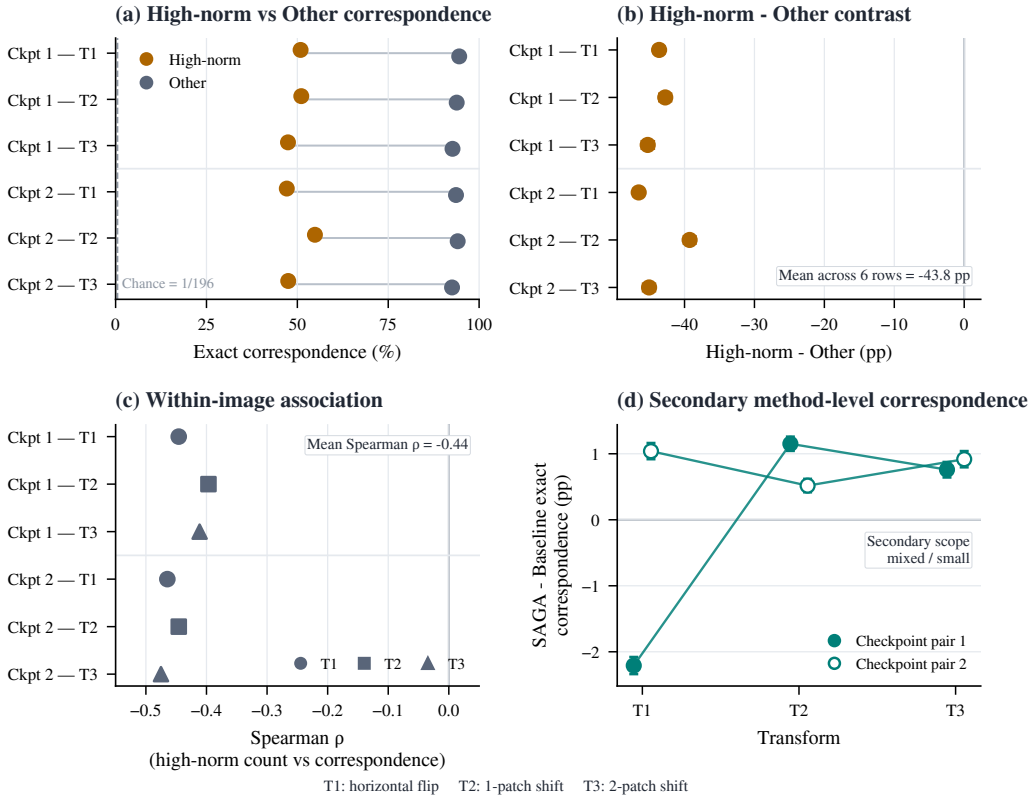

Figure 11: **Complete exact-transform patch-correspondence analysis at 's11_out'.** (a) Correspondence for high-norm and other patch positions across two fresh ViT-S/Mix Baseline checkpoints and three frozen transforms: horizontal flip (T1), one-patch shift (T2), and two-patch shift (T3). High-norm positions achieve approximately 47–55% exact correspondence, compared with 93–95% for other positions; chance is $1/196 = 0.51\%$. (b) High-norm-minus-other correspondence for all six checkpoint–transform rows. All differences are negative, with a mean of -43.8 percentage points, and all corresponding image-bootstrap intervals exclude zero. (c) Within each checkpoint and transform, images with more high-norm positions also tend to have poorer correspondence; the mean Spearman association across the six rows is $\rho = -0.44$. (d) The secondary SAGA-minus-Baseline correspondence comparison is mixed for T1 and small for T2/T3, and is shown only to delimit the method-level scope of the result.

retain 93–95%. Averaged over the six checkpoint–transform rows, the paired difference is -43.8 percentage points. The within-image analysis gives a complementary result: images containing more high-norm positions tend to exhibit poorer correspondence, with mean Spearman $\rho = -0.440$ across the same six rows.

These results give the high-norm diagnostic a nontrivial patch-consumer interpretation, but the boundary is important. The correspondence analysis is associative: it does not establish that high norm itself causes the loss of local correspondence, nor that exact geometric correspondence is a universal measure of representation quality. Likewise, the secondary method-level comparison in Figure 11(d) does not show a consistent SAGA advantage. Together with the exact terminal control, the result instead demonstrates that the meaning of a representation diagnostic depends on the consumer that reads the representation.

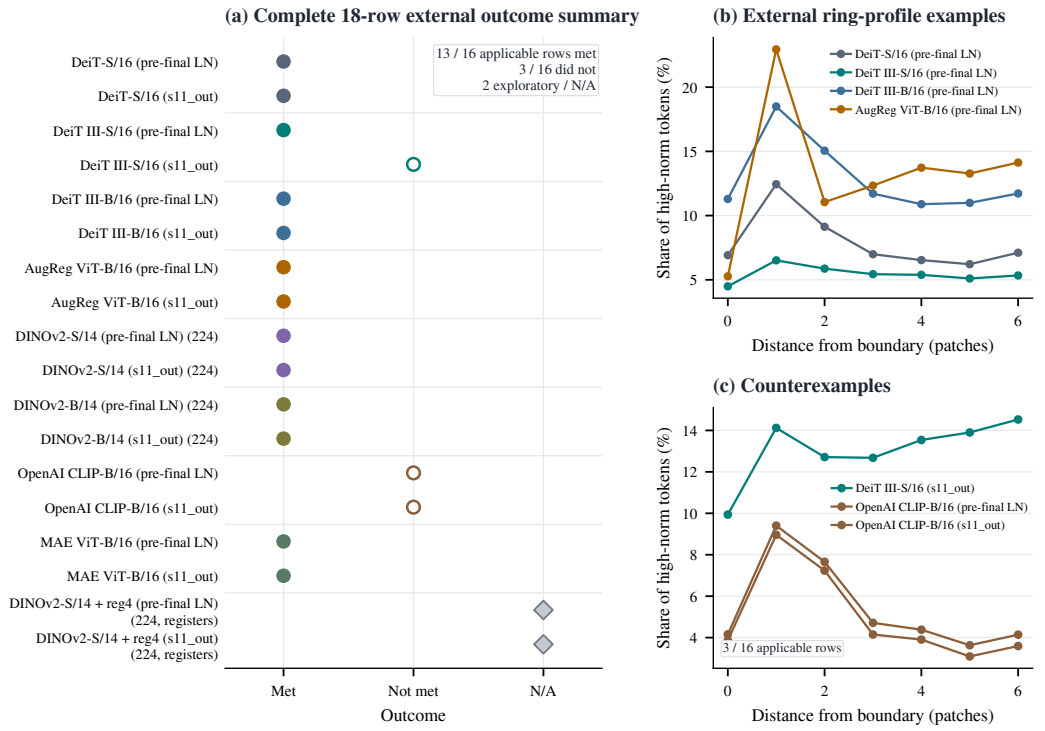

Figure 12: **External-model outcome summary and representative profiles. (a)** Registered outcomes for all 18 model–stage rows. Thirteen of the 16 applicable rows meet the declared spatial prediction, three do not, and two exploratory DINOv2-register rows are marked N/A. **(b)** Representative boundary-ring profiles from external rows that meet the prediction. **(c)** The complete set of three counterexamples: DeiT III-S/16 at 's11_out' and OpenAI CLIP-B/16 at both evaluated stages. Overall, the external audit supports partial, rather than universal, scope for the observed spatial organization.

E External-Model Scope

The main spatial analyses focus on the supervised ViTs in our trained cohort. To assess bounded external scope, we also audit a set of public pretrained checkpoints spanning DeiT (Touvron et al., 2021a), DeiT III (Touvron et al., 2022), AugReg (Steiner et al., 2021), DINOv2 (Oquab et al., 2024), OpenAI CLIP (Radford et al., 2021), and MAE (He et al., 2022). This audit contains 18 model–stage rows. Two exploratory DINOv2-register rows are retained for completeness but marked N/A and excluded from the applicable denominator.

Among the 16 applicable rows, 13 satisfy the declared prediction and three provide explicit counterexamples (Fig. 12a). The failures are not confined to a single visualization choice: DeiT III-S/16 at 's11_out' and both evaluated OpenAI CLIP-B/16 stages do not show the same ring-profile outcome. We retain these rows explicitly, rather than summarizing only the models that agree with the supervised cohort.

The complete atlas in Fig. 13 shows that the external models vary substantially in both the strength and geometry of their spatial frequency maps. For DINOv2, evaluation at 224 resolution produces a 16×16 patch grid rather than the 14×14 geometry of the patch-16 models, and we preserve that native geometry rather than resampling the maps. The two DINOv2-register rows remain exploratory/N/A and are not used to increase either the numerator or denominator of the registered outcome.

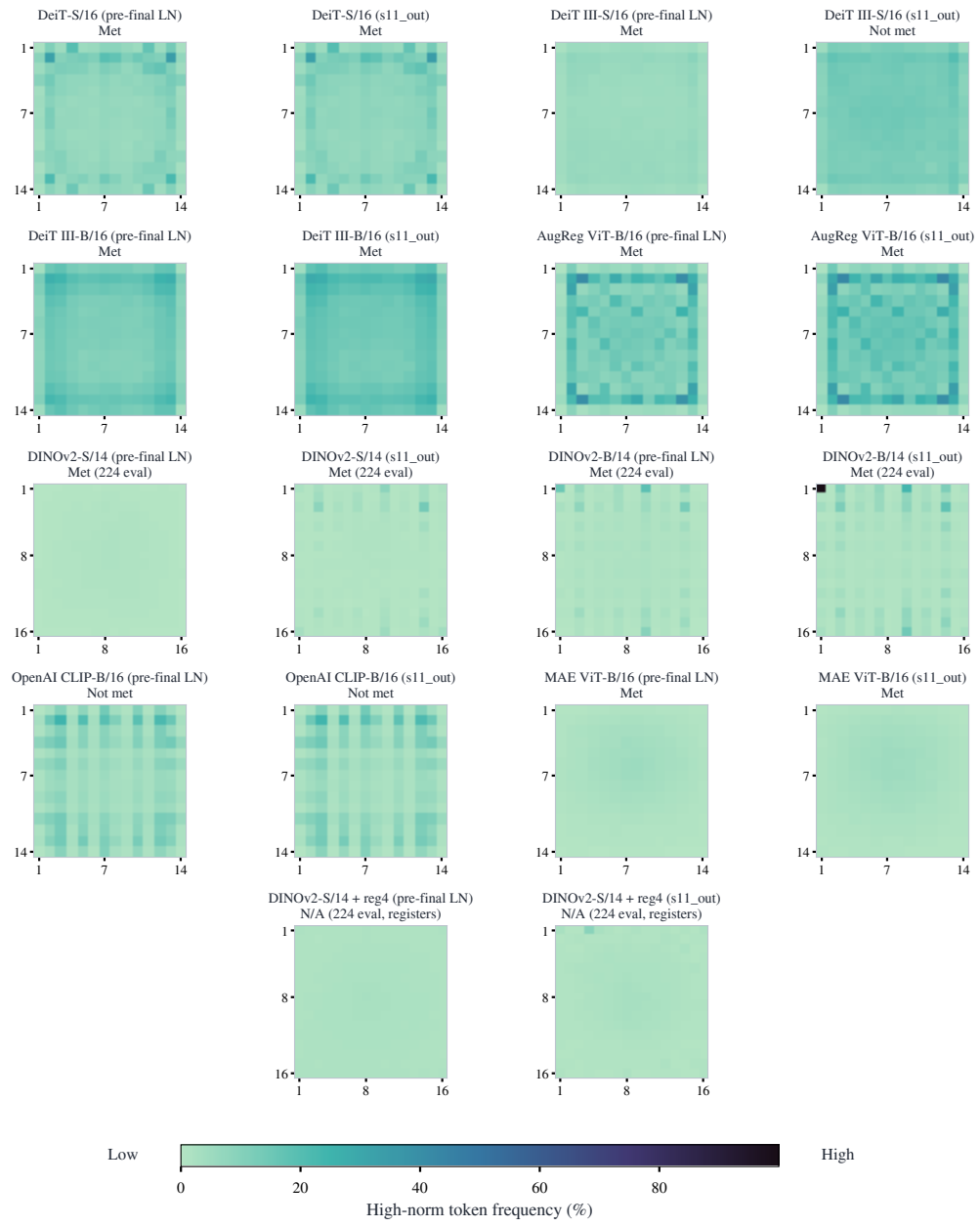

Figure 13: **Complete external spatial atlas.** Absolute high-norm-token-frequency maps for all 18 external model–stage rows, shown on a shared color scale. Native patch-grid geometry is preserved, including the 16×16 grids of the DINOv2 models evaluated at 224 resolution. Labels indicate whether each row met the declared prediction, did not meet it, or was treated as exploratory/N/A. The atlas shows that external models differ substantially in the strength and geometry of their spatial frequency maps.

Overall, this external analysis shows that the spatial pattern observed in our supervised cohort also appears in several distinct public pretrained models, but it is not universal. Because architecture,

Table 4: **Complete 300-epoch ImageNet-1K Baseline–SAGA comparison.** Top-1 values are cell means. "Paired Δ top-1" lists all SAGA-minus-Baseline checkpoint differences. Fixed is the baseline-calibrated threshold count; MAD is the per-image median $+5\,\mathrm{MAD}$ count.

Cell	n	Top-1 B	Top-1 S	Paired Δ top-1 (pp)	Fixed B→S	MAD B→S
ViT-S / Mix	4	78.720	79.176	$-0.126,\ +0.570,\ +0.490,\ +0.890$	16.480→8.401	16.276→14.435
ViT-S / NoMix	2	73.035	73.362	$+0.334,\ +0.320$	4.051→1.433	3.094→2.822
ViT-B / Mix	2	76.925	77.363	$+0.724,\ +0.152$	8.404→4.534	9.921→16.929

training procedure, and evaluation geometry all vary across these checkpoints, we treat this section as scope evidence only and do not attribute the pattern to any single training mechanism.

F Attention-Sink Verification and Test-Time Registers

The main paper verifies that high-norm patch positions receive substantially more incoming patch-query attention than other positions in the fresh ViT-S/Mix Baselines. Across the two checkpoints and Blocks 10–11, the high-norm-to-other incoming-attention ratio ranges from $6.15\times$ to $13.75\times$, with the smallest 95% confidence-interval lower bound still above $6\times$. This supports the use of *attention sink* for the measured high-norm positions, while remaining an association at the declared layers rather than a causal explanation of their formation.

F.1 Test-Time Register Operating Curves

Test-Time Registers (TTR) provide a complementary inference-time comparison. Unlike SAGA and the trained-register models, TTR does not constitute an additional trained condition: it modifies an existing Baseline only during the forward pass. We therefore evaluate TTR as an operating curve within each source checkpoint rather than as another independently trained model.

The operating curves show that high-norm-token suppression and classification accuracy do not follow a single common ordering. For the fresh ViT-S/Mix reference, one stored TTR condition changes the high-norm-token count from 19.677 to 1.912 per image while top-1 changes from 78.862% to 78.650% (-0.212 pp). The corresponding legacy ViT-S/Mix example changes the count from 15.391 to 0.973 while top-1 changes from 78.876% to 78.684% (-0.192 pp).

The same degree of classification stability does not hold uniformly across source checkpoints. For the displayed ViT-B/Mix operating point, high-norm tokens decrease from 11.286 to 5.054 while top-1 decreases by 0.976 pp. In ViT-S/NoMix, the corresponding displayed operating point changes the count from 3.671 to 2.148 with a substantially smaller 0.058 pp reduction in top-1. The complete sweep in Fig. 14 therefore reinforces the distinction between diagnostic suppression and classifier behavior: a larger reduction in high-norm-token prevalence does not by itself determine the associated accuracy change.

These TTR results are not used to rank TTR, trained Registers and SAGA as interchangeable methods. They correspond to different interventions and, in the case of TTR, are derived from an already trained source model. Their role here is to provide an inference-time example in which high-norm-token removal and classifier response must be reported separately.

G Complete Trained-System and Downstream Results

This section collects the trained-system results summarized in the main paper. The 300-epoch comparisons use the audited cohort defined in Appendix A. The 100-epoch control family is a separate single-seed ablation reported as supporting evidence rather than a replicated comparison.

ImageNet trained-system results. Table 4 reports the complete paired top-1 differences together with both the fixed-threshold and relative per-image $+5\,\mathrm{MAD}$ diagnostics. The fixed-threshold count decreases for SAGA in all three cells, whereas the relative diagnostic does not move uniformly: it decreases for both ViT-S cells but increases for ViT-B/Mix.

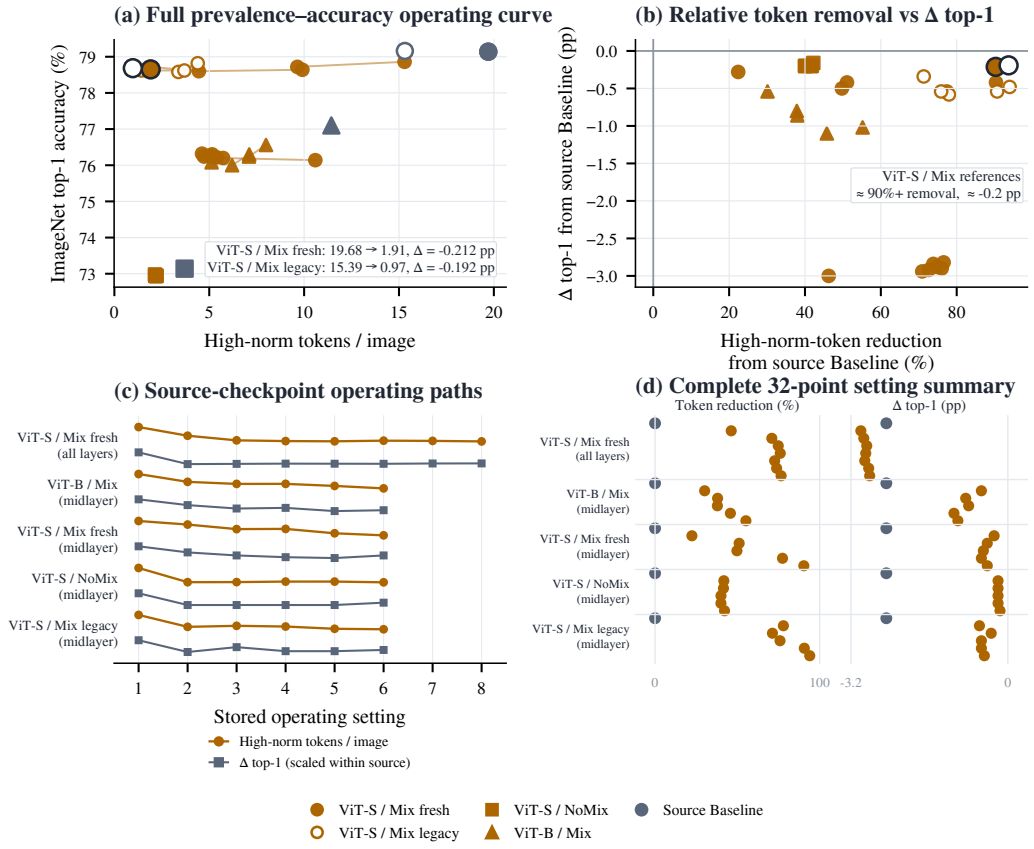

Figure 14: **Complete Test-Time Register operating sweep.** **(a)** ImageNet-1K top-1 accuracy versus high-norm tokens per image for the stored TTR settings and their source Baselines. Two ViT-S/Mix operating points remove approximately 90% or more of the source high-norm tokens while reducing top-1 by about 0.2 percentage points. **(b)** The same trade-off expressed as relative high-norm-token reduction and change in top-1 from the corresponding source Baseline. **(c)** Complete operating paths within each source checkpoint. The displayed ΔTop-1 trajectory is scaled within each source checkpoint only to visualize its path alongside token prevalence; unscaled accuracy changes are shown in panels (a), (b), and (d). Stored-setting indices identify the recorded operating configurations and should not be interpreted as training progression or as a common continuous parameter shared across checkpoint families. **(d)** Complete 32-setting summary of relative token reduction and raw ΔTop-1. TTR is an inference-time edit of the corresponding Baseline, not an independently trained model.

Scaling and initialization controls. Table 5 gives the existing six-arm 100-epoch ViT-S/Mix ablation. The baseline-calibrated fixed threshold is saturated in this short-training family, so the relative $+5$ MAD diagnostic is reported instead. Each arm is a single training run and should therefore be read as a control study rather than a replicated ranking.

Arm D is an all-token pre-projection channel-scaling control rather than a faithful implementation of canonical LayerScale. Likewise, Arm F is not a clean initialization-only intervention: $\phi = 4$ gives $\sigma(\phi) \approx 0.982$ but also changes the initial forward scale and sigmoid-gradient regime. The ablation therefore limits the explanation that the headline result follows solely from a fixed 0.5 attenuation, without establishing that spatial gating is uniquely necessary.

Downstream evaluations. Table 6 reports all evaluated fine-grained and dense-prediction outcomes. Fine-grained values are means over three paired fine-tuning seeds; COCO and ADE20K are

Table 5: **Single-seed 100-epoch scaling and initialization controls.** Δ top-1 is relative to the ungated Baseline.

Arm	Control	Top-1	Δ	+5 MAD
A	Ungated Baseline	76.526	0.000	11.713
B	Frozen gate 0.5	76.238	−0.288	4.433
C	Learned per-head scalar	76.178	−0.348	6.656
D	All-token channel scale	75.982	−0.544	17.458
E	Spatial, $\phi = 0$	76.462	−0.064	4.685
F	Spatial, $\phi = 4$	76.702	+0.176	10.026

Table 6: **Complete downstream summary.** Fine-grained Δ values are paired mean±SE over three fine-tuning seeds. COCO and ADE20K are single-run evaluations.

Dataset	Backbone	Metric	Baseline	SAGA	Δ
CUB	ViT-S/16	Top-1	80.877	80.515	-0.362 ± 0.115
CUB	ViT-B/16	Top-1	80.336	81.383	$+1.047 \pm 0.513$
FGVC-Aircraft	ViT-S/16	Top-1	73.067	74.027	$+0.960 \pm 0.442$
FGVC-Aircraft	ViT-B/16	Top-1	72.527	73.617	$+1.090 \pm 0.416$
COCO	ViT-B/16	AP	34.912	35.339	+0.427
COCO	ViT-B/16	AP_S	18.460	17.616	−0.844
ADE20K	ViT-B/16	$mIoU_{ss}$	43.577	43.335	−0.242
ADE20K	ViT-B/16	$mIoU_{ms}$	43.966	43.969	+0.002

single-run comparisons. Results are mixed rather than uniformly positive: three of four fine-grained cells favor SAGA, COCO overall AP increases while AP_S decreases, and ADE20K is unchanged under multi-scale evaluation but slightly lower under single-scale evaluation.