READ WHERE THE MODEL READS: CAUSAL ANATOMY OF VIT CLASSIFICATION AND ITS INSTRUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In language models, the logit lens interprets intermediate representations by applying the final unembedding directly to the current token state. This is natural because the current token stream ultimately reaches the unembedding. However, in Vision Transformers (ViTs), the classifier reads only the CLS stream, while image information resides in the patch tokens. Despite this structural difference, direct projection of intermediate representations through the final classifier is routinely used to interpret ViTs. We show that this routine can be misleading. Across five ViT backbones, we first examine the causal anatomy of classification and find that class evidence develops within the patch stream and is written into CLS only through a few terminal layers, which we call the *read window*. Motivated by this anatomy, we introduce the read-through lens (RTLens), a logit lens variant that passes intermediate patch states through the read window before applying the final classifier. We further derive an anatomy-based attention map that exactly decomposes patch contributions to the prediction through the read window. Across all five backbones, RTLens predictions match the model's causal behavior far better than direct projection, and our map more accurately localizes the evidence the model actually uses. Together, these results support a simple principle for interpreting ViTs, *read where the model reads*.

1 INTRODUCTION

In language models, intermediate residual states and the final readout share a common tokenwise geometry. The unembedding acts on the residual stream at each token position, so projecting an intermediate state into vocabulary space is a natural diagnostic. The logit lens (nostalgebraist, 2020; Belrose et al., 2023) and related inspection methods build on this geometry, while circuit analyses test the causal role of the decoded content through intervention (Elhage et al., 2021; Wang et al., 2022). However, Vision Transformers (ViTs) differ structurally. A ViT classifier reads only the final CLS state, while patch tokens carry the spatially distributed image representation and are never read directly by the classifier (Dosovitskiy, 2020). Class-relevant information must therefore cross token positions before reaching the classifier. Yet, logit-lens projections are routinely applied to intermediate ViT states (Vilas et al., 2023; Neo et al., 2025), and attention-based attribution methods often aggregate or propagate information across layers without causally identifying which layers deliver evidence to the classifier (Abnar & Zuidema, 2020; Chefer et al., 2021a).[1]

To test whether these interpretability routines can be transplanted directly to ViTs, we first examine the model's reading geometry. Using distribution-preserving interventions (Chan et al., 2022) across five ViT backbones, we screen every feasible attention route among CLS, register, and patch tokens and test whether intermediate states are causally used. Across models, the interventions reveal a consistent anatomy (Fig. 1a). Patch-to-patch interactions carry the dominant cross-token computation through the early and middle layers, while patch-to-CLS attention transfers class evidence into CLS only within a narrow terminal *read window*.

These results reveal a structural limitation of directly transferring decoding-based interpretation from language models to ViTs. A direct projection through the final readout measures whether an intermediate state is aligned with that readout, but does not establish whether downstream computation

[1] An extended discussion of related work is provided in §B.

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.

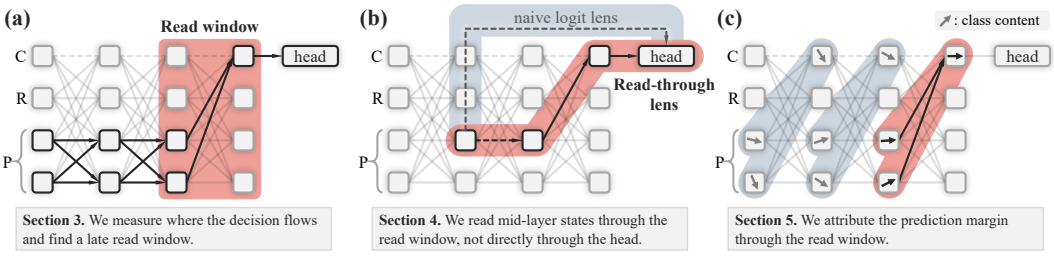


Figure 1: Overview of the paper. Red marks anatomy-derived structures and designs, while blue-gray marks existing interpretability practices that do not account for the measured ViT anatomy. **(a)** Class evidence forms in the patch stream and reaches CLS only through a terminal read window, while CLS before the read window acts as a placeholder (§3). **(b)** RTLens reads patch states through the read window rather than directly through the final head (§4). **(c)** Our attention map attributes the prediction margin only to messages delivered through that window rather than across all layers (§5).

actually uses it. CLS states before the read window make this distinction concrete: although logit-lens methods can already recover the image class from these states, replacing them with CLS states from another image barely changes the model's prediction. Decodability therefore does not indicate causal use. We thus adopt a simple principle for building interpretability instruments: *read where the model reads*.

We apply this principle to decoding to devise the *read-through lens* (RTLens, §4). RTLens transports an intermediate patch state through the read window before applying the frozen classifier, aligning the projection with the path by which patch information reaches the classifier. Neuron-level clamp interventions show that RTLens substantially improves causal agreement over the naive logit lens. As an external check, we further show that an averaged Jacobian lens (Gurnee et al., 2026) approaches a similar transport near the read window.

We next apply the same principle to attribution to derive an attention map for ViTs (§5). Direct logit attribution (DLA) decomposes a prediction into additive contributions delivered through attention, which prior work evaluates at the final layer or across all layers (Elhage et al., 2021; Gandelsman et al., 2024a). We keep this accounting but evaluate it only over the read window, so each patch is scored by the content it actually delivers to the classifier. The resulting map is an exact, gradient-free decomposition of the patch-delivered component of the prediction margin. Experiments show that our map achieves the highest energy localization among the compared methods and that its high-scoring patches are generally more causally selective than those of the all-layer DLA map.

Our contributions are summarized as follows:

- **(Causal anatomy)** A two-stage anatomy of patch-stream computation and a late read window.
- **(Decodable $\neq$ computed)** CLS states before the read window are decodable yet causally near-inert.
- **(Read-through lens)** A causally validated lens reading patch states through the read window.
- **(Attention map)** A better-localized, causally verified map scoped to the read window.

2 Setup

Models and data. We evaluate five ViT-B backbones spanning self-supervised, contrastive, and supervised training: DINOv2 ViT-B/14 with four register tokens (Oquab et al., 2023; Darcet et al., 2024), DINOv3 ViT-B/16 with four register tokens and RoPE (Siméoni et al., 2025), CLIP ViT-B/16 (Radford et al., 2021), supervised ViT ViT-B/16 from the AugReg lineage (Dosovitskiy, 2020; Steiner et al., 2021), and DeiT-B/16 (Touvron et al., 2021), together with DINOv2 ViT-L/14 as a scale probe. All experiments use ImageNet-1k at 224^2 resolution.

Readouts. The supervised ViT and DeiT models come with native classification heads, while CLIP provides a zero-shot classifier through its text encoder. The self-supervised DINOv2 and DINOv3 backbones do not have native ImageNet classification heads, so following standard evaluation practice (Oquab et al., 2023; Siméoni et al., 2025), we train a linear probe on their frozen final-norm CLS representations. From a fixed shuffle of the 50,000 validation images, 20,000 are reserved for probe

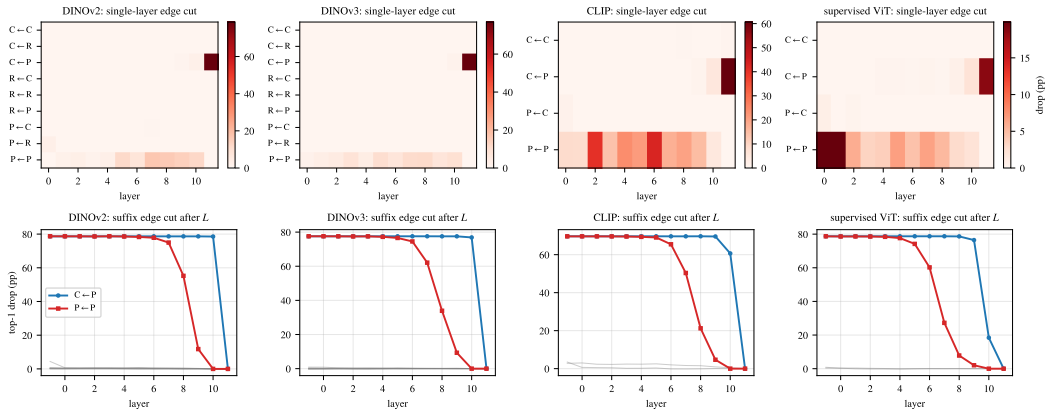


Figure 2: Causal screen of all feasible token-group attention edges. **(Top)** Single-layer edge resampling. **(Bottom)** Suffix interventions that block an edge at every layer after L. Gray curves denote the remaining edges other than $C \leftarrow P$ and $P \leftarrow P$.

training, and disjoint evaluation samples are drawn from the remaining 30,000. Claims for the self-supervised backbones therefore apply to the model–probe pair. Readout and probe details, including the readouts used for each experiment, are provided in §C.

Statistics. For reproducibility, all experiments use fixed random seeds and two-sided tests without multiplicity correction. Uncertainty is reported as paired bootstrap confidence intervals (CIs). Sample sizes are chosen per experiment based on intervention cost and stated where each experiment is introduced. Complete sample sizes and donor assignments are given in §C.

Notation. We use C, R, and P to denote CLS, register, and patch tokens, respectively, and write $\leftarrow$ for an attention edge. For example, $C \leftarrow R$ denotes an attention edge from register tokens to CLS. Layers are zero-indexed and written Lℓ, so L1 denotes the second layer of a ViT-B backbone.

3 The Causal Anatomy of ViT Classification

We first identify the token interactions that carry the classification decision and then test whether intermediate CLS content is causally used. The results reveal a two-stage mechanism: ① class evidence is formed within the patch stream, and ② then transferred to CLS through a late read window (§3.1). Before this transfer, CLS primarily functions as a content-light placeholder (§3.2).

3.1 A Two-Stage Anatomy

Two attention routes dominate the causal effect. To identify which attention routes causally support classification, we ablate each feasible attention edge among the C, R, and P token groups. Following recent activation-patching practice (Heimersheim & Nanda, 2024; Zhang & Nanda, 2024), we use *resampling ablation*, which replaces the contribution along an edge with the corresponding contribution from another image rather than deleting it, keeping the intervention closer to the model's activation distribution (Vig et al., 2020; Meng et al., 2022; Conmy et al., 2023; Chan et al., 2022). We use two resampling schedules and measure the resulting top-1 drop (n=4,000 images): a *single-layer* intervention, which resamples one edge at one layer, and a *suffix* intervention, which resamples the same edge at every layer after L. The results in Fig. 2 show that two routes carry most of the causal effect across all backbones. Mid-layer $P \leftarrow P$ supports computation within the patch stream, while late $C \leftarrow P$ transfers patch information to CLS (Fig. 3). The remaining edges have consistently small effects under both schedules. The suffix intervention further supports this picture by showing that

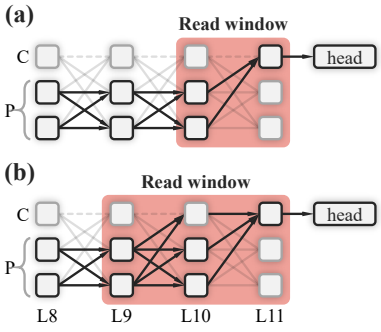


Figure 3: **(a)** Single-read and **(b)** dual-read anatomy.

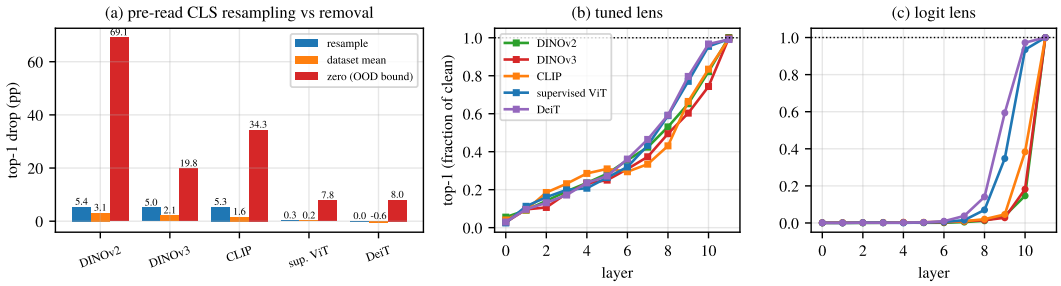


Figure 4: Pre-read CLS content is decodable yet largely unused. **(a)** Top-1 drops from replacing pre-read CLS states with another image's states, a dataset-mean state, or zeros. **(b)** Tuned-lens and **(c)** logit-lens trajectories across backbones, each normalized by its clean accuracy.

the effect of $P \leftarrow P$ is distributed across multiple layers. We refer to this pattern, mid-layer $P \leftarrow P$ computation followed by a late $C \leftarrow P$ read, as the *two-stage anatomy*.[2]

Supervised models use a wider terminal read. We next ask whether the terminal $C \leftarrow P$ read is concentrated in the final layer or spread across the final two layers. To test this, we compare the joint L10–11 intervention with the stronger of the two single-layer interventions. As shown in Table 1, DINOv2, DINOv3, and CLIP are dominated by L11, with little additional effect from intervening on L10 and L11 together. In contrast, supervised ViT and DeiT are more strongly affected when both layers are intervened on jointly, indicating that both L10 and L11 contribute to the read. We therefore define the *read window* $\mathcal{W}$ as L11 for DINOv2, DINOv3, and CLIP (Fig. 3a), and L10–11 for the two supervised models (Fig. 3b). §F provides the formal criterion, measurement details, and analysis at the ViT-L scale.

Table 1: Top-1 accuracy drops (pp) from resampling the $C \leftarrow P$ edge at the listed layers, with clean top-1 accuracy (%) for reference.

	$C \leftarrow P$ resampling			
Model	L10	L11	L10–11	clean (%)
DINOv2	2.3	78.6	78.7	79.5
DINOv3	1.5	76.9	77.5	78.2
CLIP	1.7	63.8	66.6	67.2
sup. ViT	1.4	16.4	79.8	81.7
DeiT	0.7	3.0	68.6	82.0

3.2 Pre-read CLS is a content-light placeholder

Image-specific pre-read CLS content has little causal effect. We next test whether the class information decodable from pre-read CLS (*i.e.*, CLS states before the read window) is causally used. At every layer below the read window $\mathcal{W}$, we jointly *replace* the CLS states either with the corresponding states from another image or with a fixed per-layer dataset mean. We also *zero* the same states as an out-of-distribution control on token presence (Miller et al., 2024; Li & Janson, 2024). Fig. 4a shows a consistent pattern across backbones. Replacing CLS content has only a small effect on accuracy, whereas zeroing is much more damaging. Before the read window, the model therefore depends on an in-distribution CLS state but only weakly on which image-specific content that state carries. We call this behavior a *content-light placeholder* (precise criterion in §F). §D further shows that replacing pre-read CLS states with Gaussian noise matched to their per-channel statistics is similarly harmless. Note that our result extends the concurrent register-token findings of Parodi et al. (2026) to the CLS token, with their mechanistic differences detailed in §E.

Decoding nonetheless recovers class information from these states. We decode the same pre-read CLS states with the *logit lens*, which applies the frozen final readout directly (nostalgebraist, 2020; Vilas et al., 2023), and the learned *tuned lens* (Belrose et al., 2023). As shown in Fig. 4b,c, the tuned lens recovers class information well before the read window across all backbones, while the naive logit lens becomes decodable below the read window in the supervised models. Thus, pre-read CLS states can be strongly decodable even when their image-specific content has little causal effect. Decodability alone therefore does not identify where the model actually uses that information. This extends a known caution from language-model probing (Elazar et al., 2021) to ViT classification.

[2]The two-stage anatomy describes the principal attention routes rather than the complete classification circuit, which also includes within-token transformations such as MLPs (extended discussion in §H).

The same dissociation persists under native readouts (§G.5) and at the ViT-L scale (§F.3), motivating the principle developed in the next two sections: **read where the model reads**.

4 THE LOGIT LENS FOR VITS

§3 showed that ViTs form class evidence in the patch stream and transfer it to CLS through a late read window. We therefore propose the read-through lens (RTLens), a logit-lens variant that reflects the operation of the read layers (§4.1). We then validate its readings causally (§4.2) and use RTLens to interpret the patch stream (§4.3).

4.1 RTLENS FOLLOWS THE MEASURED CAUSAL PATH

The naive logit lens applies the final readout directly to an intermediate state. For patch states, this projection does not match the model's computation (§3.2), since patch content reaches the classifier only after passing through the read window. We therefore define RTLens, which applies this read operation before the final classifier.

Let $P^{(\ell)} \in \mathbb{R}^{N\times d}$ denote the patch states at the output of layer ℓ, with $P_n^{(\ell)} \in \mathbb{R}^d$ denoting the patch n. Here, N is the number of patches and d is the residual width. Using the read window $\mathcal{W}$ identified in §3.1, we define $\mathcal{W}_\ell = \{r \in \mathcal{W} \mid r > \ell\}$ as the read layers that remain above layer ℓ. In particular, $\mathcal{W}_{11} = \varnothing$ at the final layer, so no patch readout is defined there. For each $r \in \mathcal{W}_\ell$, we evaluate the message that read layer r delivers from patch n to CLS using $P_n^{(\ell)}$ as the source state:

$$\mathrm{msg}^r\big(P_n^{(\ell)}\big) \coloneqq \sum_h A^{r,h}[\mathrm{CLS}, n]\; W_O^{r,h} W_V^{r,h}\; \mathrm{LN}_1^r\big(P_n^{(\ell)}\big), \tag{1}$$

where $A^{r,h}[\mathrm{CLS}, n]$ is the realized attention weight from CLS to patch n in head h of read layer r, $W_V^{r,h}$ and $W_O^{r,h}$ are the corresponding value and output projections, and LN_1^r is the pre-attention LayerNorm.[3]

RTLens computes the patch-to-CLS content that the remaining read window would deliver from the layer-ℓ patch states and then applies the frozen readout. Writing $z_c : \mathbb{R}^d \to \mathbb{R}$ for the logit of the c^{th} class, we define RTLens for both the whole patch stream and an individual patch as

$$\text{(Whole patch stream)} \qquad \mathrm{RTLens}\big(P^{(\ell)}\big) \coloneqq \arg\max_c z_c\Big(\sum_{r\in\mathcal{W}_\ell}\sum_n \mathrm{msg}^r(P_n^{(\ell)})\Big), \tag{2}$$

$$\text{(Single patch)} \qquad \mathrm{RTLens}\big(P_n^{(\ell)}\big) \coloneqq \arg\max_c z_c\Big(\sum_{r\in\mathcal{W}_\ell} \mathrm{msg}^r(P_n^{(\ell)})\Big). \tag{3}$$

Lenses are also used to interpret individual neurons by projecting their output directions through the readout (Geva et al., 2022; Dar et al., 2023; Gandelsman et al., 2024b). We therefore define RTLens on neuron write directions. For a neuron contributing aw to the residual stream, where $a \in \mathbb{R}$ is its activation, we call $w \in \mathbb{R}^d$ its write direction. Since w is a fixed write direction rather than an image-dependent activation, its label should not depend on any particular image. Following weight-space analyses of the value–output circuit, which treat attention as a scalar coefficient (Elhage et al., 2021; Dar et al., 2023), we remove the image-dependent attention weights and per-token normalization scale from Eq. (1), retaining only the direction-level value–output transport. Thus, the RTLens for a write direction is

$$\text{(Neuron write direction)} \qquad \mathrm{RTLens}(w) \coloneqq \arg\max_c z_c\Big(\sum_{r\in\mathcal{W}_\ell}\sum_h W_O^{r,h} W_V^{r,h}(\gamma_1^r \odot (w - \bar{w}\mathbf{1}))\Big), \tag{4}$$

where γ_1^r is read layer r's LayerNorm gain and $\bar{w}$ is the mean of w (§G, Eq. 10). Note that RTLens requires neither additional parameters nor training.

4.2 CAUSAL VALIDATION AND LAYERWISE TRAJECTORIES

We validate RTLens against causal interventions at two levels, corresponding to its two forms. At the neuron level, we test whether the label assigned by Eq. (4) predicts the class causally promoted when the neuron is activated. At the state level, we apply Eq. (2) across layers and measure how much class information the executed read recovers from intermediate patch states.

[3]For readability, the main-text formulas omit attention-block bias terms and LayerScale gains. All experiments and exactness claims use the complete operator, whose full form is given in §G.

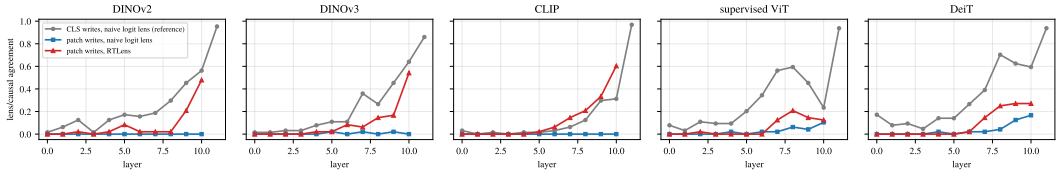


Figure 5: Patch-stream clamp test across five backbones. Agreement measures whether the class causally promoted by clamping a neuron matches the class assigned to its write direction by each lens. The CLS reference profile is shown in gray, the naive logit lens in blue, and RTLens in red.

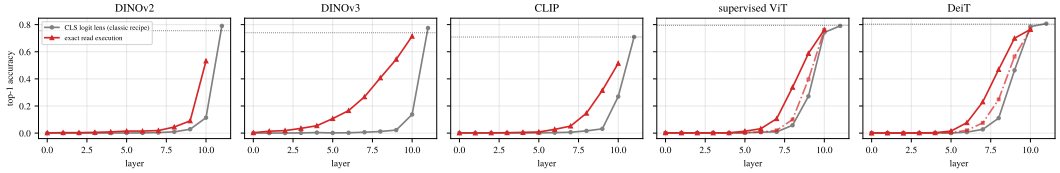


Figure 6: Layerwise decoding from intermediate patch states with n=2,000 images. The CLS logit lens is shown in gray, RTLens over the full read window in solid red, and the terminal read layer alone in dash-dotted red. The dotted line marks final accuracy.

Read-through labels align with causal effects near the read window. We randomly sample 48 MLP neurons per layer and clamp each neuron individually to its 99th-percentile activation at all patch positions, following activation-clamping interventions used to verify unit semantics in language models (Dai et al., 2022; Templeton et al., 2026). For each neuron, we compare the class most promoted by the intervention with the class assigned to its write direction by the naive logit lens and RTLens. Because both lenses use the same neurons and interventions, differences in agreement isolate the projection. As shown in Fig. 5, the naive logit lens never exceeds 0.17 agreement across the five backbones, whereas RTLens reaches substantially higher agreement near the read window, including 0.60 on CLIP at L10, well above the empirical label-frequency null (§G). Agreement for both lenses remains low at earlier layers, where patch writes undergo additional $P \leftarrow P$ interactions before reaching the read window.

Agreement at CLS does not transfer to the patch stream. The gray curve in Fig. 5 applies the same clamp protocol at the CLS position, labeling each write direction with the naive logit lens. Its agreement is generally higher than at patch positions, since a write inserted at CLS reaches the classifier more directly through the residual stream.[4] The contrast with the patch-position curves is the important control. When the same naive labeling is applied to writes inserted at patch positions, agreement drops sharply because those writes must still pass through the terminal read before reaching the classifier. RTLens improves agreement by incorporating that transport. Thus, the CLS result shows that the clamp test can recover a neuron's causal label when the projection is well matched to the route by which its write reaches the classifier.

Read execution reveals class evidence earlier in the patch stream. We next apply Eq. (2) across layers to ask how much class information can be recovered from intermediate patch states once they are read through the measured path. As shown in Fig. 6, RTLens outperforms the naive logit lens across all five backbones from L4 onward below the read window, while both remain near chance at earlier layers. Since RTLens introduces no learned parameters, this improvement comes from executing the measured read rather than learning a layer-specific correction. In the dual-read models, using the full read window consistently outperforms using the terminal layer alone, in agreement with the two-layer read identified in §3.1. §G reports the full per-model results and representative patch-label maps.

An independently derived lens aligns with the same transport. As an external check, we port the Jacobian lens (JLens) of Gurnee et al. (2026) to three backbones. JLens linearizes the forward map from an intermediate layer to the final representation by averaging its Jacobian over images, without using the causal anatomy. We compare each stream's averaged Jacobian with the read-through

[4]This does not contradict §3.2: CLS replacement measures effect magnitude, whereas the clamp test asks whether a synthetic write's causal direction is labeled correctly.

Table 2: Operator geometry of the averaged Jacobian. Matrix cosine similarity of the patch- and CLS-position Jacobians to the read-through transport and to the identity transport that direct projection by the naive logit lens implies. Shaded rows indicate the anatomy-consistent transport for each stream: read-through for patches and identity for CLS.

		Source layer L9			Source layer L10		
Jacobian at	Compared with	DINOv2	sup. ViT	DeiT	DINOv2	sup. ViT	DeiT
Patch	read-through (RTLens)	0.58	0.67	0.78	0.85	0.96	0.97
Patch	identity (naive)	-0.04	0.33	0.28	0.03	0.39	0.35
CLS	read-through (RTLens)	-0.01	0.40	0.35	0.02	0.40	0.35
CLS	identity (naive)	0.77	0.98	0.97	0.86	0.99	0.99

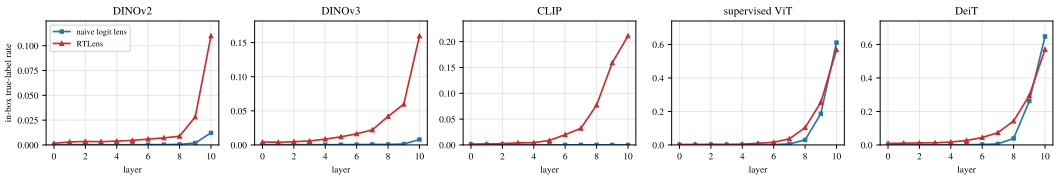


Figure 7: Per-patch class labels across layers. Curves show the fraction of bounding-box patches assigned the true class on finally correct images (n=500 per model). Blue denotes the naive logit lens and red the read-through projection.

transport and the identity transport corresponding to direct projection by the naive lens. As shown in Table 2, the patch Jacobian aligns more closely with read-through near the read window, whereas the CLS Jacobian remains close to identity. Although some patch alignment is expected from the terminal value-output pathway, note that it is not guaranteed by the full downstream Jacobian. These results are consistent with patch information reaching the classifier through the terminal read, while a downstream-aware first-order lens at CLS can still report decodability that the model does not causally use. Protocol details are given in §G.

4.3 Interpreting the patch stream with RTLens

Per-patch class projection is a common way to interpret spatial representations in ViTs (Vilas et al., 2023; Schaffer et al., 2025). However, our causal anatomy shows that intermediate patch states do not reach the classifier directly. This suggests that patch states should be decoded only after accounting for the read through which their content reaches CLS. We therefore use RTLens to ask when and where class information becomes readable within the patch stream.

Read-through labeling reveals class information before the read window. We label each patch with the argmax class of its RTLens projection (Eq. 3) and measure the fraction of bounding-box patches assigned the image's true class (Fig. 7). On the self-supervised backbones and CLIP, the naive logit lens recovers essentially no true-class labels below the read window, whereas RTLens yields true-class labels from the same patch states. Since the two projections use the same states and frozen probe, the difference comes from whether the read operation is applied before classification. The supervised models show a different late-layer pattern. Near the output, their patch states become directly aligned with the classification head, so the naive logit lens overtakes RTLens upon entering the read window at L10, while RTLens remains ahead at every earlier layer. This path-based view also offers a way to interpret prior dense-prediction methods: the modification of Zhou et al. (2022) approximates a terminal read by retaining the value pathway while discarding its attention weights, whereas RTLens applies the realized read operation.

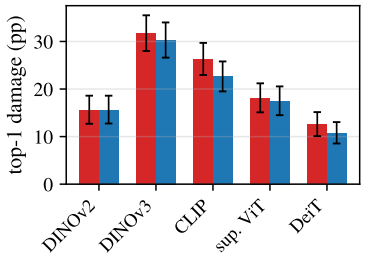
5 The attention map for ViTs

§4 introduced RTLens, which reads along the path used by the model. We apply the same principle to spatial attribution. We first construct a map from the content delivered through the read window along the classifier's decision direction (§5.1). We then test the two anatomy-derived design choices (§5.2) and verify whether high-scoring patches are causally important (§5.3).

Table 3: Energy localization on correctly predicted images with boxes (n=362−420 per model). Ours is the map s_n over the read window. The + row widens the scope one layer below the window (L10–11 and L9–11 for single- and dual-read models), and the norm-weighted row follows Kobayashi et al. (2020).

Energy pointing (↑)	DINOv2	DINOv3	CLIP	sup. ViT	DeiT
Ours ($A \times$ content)	0.887	0.917	0.859	0.865	0.864
− content (A only)	0.841	0.848	0.642	0.694	0.597
− layer scope (all layers)	0.876	0.893	0.825	0.857	0.845
+ one layer below $\mathcal{W}$	0.885	0.902	0.856	0.863	0.864
norm-weighted $A\|\|v\|\|$	0.839	0.850	0.743	0.748	0.694
Rollout	0.698	0.711	0.667	0.690	0.673
GAE (gradient)	0.814	0.822	0.756	0.774	0.760

Figure 8: Top-1 damage from replacing the top-8 patches of our map (■) and of the all-layer DLA map (■) at the read window, with paired bootstrap CIs.



5.1 An exact map of the terminal read

To read where the model reads, we start from direct logit attribution (DLA). DLA decomposes a logit into additive contributions from residual-stream writes, including attention messages through the value–output circuit (Elhage et al., 2021; Wang et al., 2022). Prior work evaluates these contributions at the final layer in language models and across all layers in CLIP (Gandelsman et al., 2024a). We retain this accounting framework but set its layer scope using the causal anatomy. Since patch content reaches the classifier through the read window, we evaluate the decomposition over $\mathcal{W}$ rather than over the final layer or all layers.

Following standard DLA practice (Wang et al., 2022), we attribute the prediction margin $z_{\hat{c}} - z_{c_2}$ between the predicted class $\hat{c}$ and the second-ranked class c_2, computed with the frozen readout z_c of §4.1. With the realized per-image LayerNorm mean and standard deviation fixed, this margin decomposes exactly as $z_{\hat{c}} - z_{c_2} = \sum_n s_n + s_{\text{glob}} + s_{\text{res}} + s_{\text{MLP}} + s_{\text{const}}$. Here, $\sum_n s_n$ is the contribution delivered from patch tokens, s_{glob} contains messages delivered from CLS and register tokens, s_{res} is the residual carried by CLS into the read window, s_{MLP} contains the CLS writes of the read-layer MLPs, and s_{const} contains the bias and intercept terms. In dual-read models, the message, MLP, and bias contributions are summed over both layers in $\mathcal{W}$.

With the LayerNorm statistics fixed, the frozen readout is affine, so the prediction margin $z_{\hat{c}} - z_{c_2}$ can be written as a projection of the final CLS state. Let q_c denote the residual-space direction of class c, obtained by absorbing the realized gain and scale of the final LayerNorm into the frozen classifier weight.[5] The margin direction is then $q = q_{\hat{c}} - q_{c_2}$, and mean-centering gives $\tilde{q} = q - \bar{q}\mathbf{1}$, accounting exactly for the centering in the final LayerNorm. Unrolling the final CLS state through the read window and distributing this projection over its additive contributions yields the per-patch terms s_n, which constitute the spatial attribution map we propose:

$$s_n = \Big\langle \sum_{r \in \mathcal{W}} \text{msg}^r(P_n),\, \tilde{q} \Big\rangle = \sum_{r \in \mathcal{W}} \sum_h A^{r,h}[\text{CLS}, n] \left\langle W_O^{r,h} W_V^{r,h}\, \text{LN}_1^r(P_n),\, \tilde{q} \right\rangle. \tag{5}$$

Each message is evaluated at the actual input to its corresponding read layer. In the second form, each patch score factors into an attention weight and a decision-aligned content term, which we compare directly in §5.2. Positive s_n indicates that the content delivered from patch n supports the predicted class over the runner-up, while negative s_n indicates support for the runner-up.

This construction follows directly from the measured causal anatomy. It attributes only the decision-aligned patch content delivered through the read window, while leaving the remaining contributions to the prediction margin separate. §H provides the full derivation and quantifies each term's contribution.

5.2 Both anatomy-derived components matter

We test the two design choices implied by the causal anatomy: the measured layer scope and the projection of delivered patch content onto the decision direction. Localization is evaluated with the

[5] Specifically, $q_c \coloneqq (\gamma/\sigma) \odot W_c$, where γ is the final LayerNorm gain, σ its realized per-image standard deviation, and W_c the class-c row of the frozen classifier. §H gives the complete form.

energy pointing game (Wang et al., 2020), which measures the fraction of positive map mass inside the object bounding box. Table 3 reports the main results, §H provides the full design grid, including prior DLA variants as individual cells.

Decision-aligned content is more informative than attention alone. Removing the content projection and retaining only the attention weights reduces energy localization across all five models, with confidence intervals excluding zero in every case. Weighting attention by the value norm $A\|v\|$ (Kobayashi et al., 2020) recovers little of this loss. Thus, the delivered content is more informative when scored by its alignment with the decision direction than by attention or value magnitude alone.

The read window provides the appropriate layer scope. Extending the same $A \times \text{content}$ score from the read window to all layers reduces localization across all five models. In the dual-read models, including both measured read layers improves localization with confidence intervals excluding zero, while the same widening does not improve any single-read model (§H). Thus, the layer scope identified by the causal anatomy also yields better localization than the corresponding all-layer DLA. Extending the scope one layer below the read window provides no further improvement in any model and significantly reduces localization on DINOv2 and DINOv3 (Table 3).

Reference methods and further comparisons. Table 3 also includes attention rollout (Abnar & Zuidema, 2020), which aggregates attention across all layers regardless of causal role, and the gradient-based GAE (Chefer et al., 2021a) as references. Both trail the read-window map on energy localization. §5.3 compares the methods under additional evaluation criteria, while §H reports alternative contrast directions, scope variants, and reproductions of prior DLA maps.

5.3 Causal verification and scope of the contribution

An exact decomposition does not by itself establish causal importance, since a patch's attributed contribution need not match the effect of perturbing that patch. We therefore test whether high-scoring patches are causally important and then compare the resulting map under standard saliency criteria.

High-scoring patches are causally selective. We select the top-8 patches from our map and the all-layer DLA map, replace their value vectors (*i.e.*, $W_V \mathrm{LN}_1(P_n)$ in Eq. 1) at the read window,[6] and measure the resulting drop in top-1 accuracy (n=500 images). As shown in Fig. 8, patches selected by our map cause greater damage than those selected by the all-layer DLA map in three of five models, with confidence intervals excluding zero. §H reports results for additional numbers of replaced patches.

Scope of the contribution. Our experiments suggest that, under matched conditions, the results favor the anatomy-derived design over its alternatives (Table 3, Fig. 8, with the full design grid in Table 10). However, comparisons across different method families yield mixed results. While our map achieves the highest localization across all five models, GAE generally leads on deletion faithfulness (Petsiuk et al., 2018) (§I). Our contribution is therefore not a uniform improvement over existing saliency methods. Rather, it lies in deriving the map's layer scope and scoring rule directly from the measured causal anatomy, yielding an exact, gradient-free decomposition of the patch-delivered component of the prediction margin.

6 Conclusion

Across five ViT backbones, we identify a common two-stage anatomy in which classification is computed in the patch stream and written into CLS by a terminal attention read, before which CLS acts as a content-light placeholder. This measured anatomy motivates a simple principle, *read where the model reads*, and yields two causally validated instruments. First, RTLens reads intermediate patch states through the read window rather than directly through the final head. Second, our attention map scores the decision-aligned content delivered by each patch within the same read window. Beyond ViT classification, our results suggest a general recipe for interpreting intermediate states: identify the causal path to the readout, determine where a projection is valid, and verify that interpretation by intervention before using it to explain or edit models. We hope this work encourages interpretability grounded in how models actually read. The limitations of this study are discussed in §A.

[6] Replacing patches at the image input does not consistently favor either map, since input-level replacement perturbs all downstream paths rather than isolating the measured read (§H).

AI USE STATEMENT

In this work, we used generative AI tools for partial translation and language polishing of the manuscript, organizing existing code, and checking consistency between the manuscript and the implementation. We did not use generative AI tools to generate research ideas, design or specify the experimental methodology, write substantive experimental code, or run experiments. Other required disclosure tasks are not applicable to this work. We reviewed all AI-assisted work. All research ideas, experimental designs and refinements, implementation decisions, experiments, and scientific conclusions were developed and carried out by the authors. AI-assisted manuscript edits were reviewed by the authors, and AI-assisted code organization and manuscript–code consistency checks were verified against the authors' implementation and experimental procedures. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We take several steps to support reproducibility. All experiments use publicly available pretrained checkpoints and public datasets, as described in §2. We use fixed random seeds throughout and report the statistical procedures used for each experiment. The complete formulations of our methods are provided in the appendix, including the read-through operators in §G, the exact margin decomposition in §H, and the formal read-window criterion in §F. §C provides dedicated reproducibility details, including the sample size of every experiment, deterministic data splits, the linear probe-training recipe, donor construction for all interventions, bootstrap granularity, and a per-model coverage table for the main-text experiments. Source code reproducing all experiments, figures, and reported numbers is submitted as supplementary material.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP:: Attention-Aware Layer-Wise relevance propagation for transformers. *arXiv preprint arXiv:2402.05602*, 2024.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31, 2018.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL `https://transformer-circuits.pub/2025/attribution-graphs/methods.html`.

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. Decomposing and interpreting image representations via text in ViTs beyond CLIP. *Advances in Neural Information Processing Systems*, 37:81046–81076, 2024.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers. *Advances in Neural Information Processing Systems*, 38:112351–112383, 2026.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Walid Bousselham, Angie Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. LeGrad:: An explainability method for vision transformers via feature formation sensitivity. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Edmund Bu and Yossi Gandelsman. Interpreting ResNet-based CLIP via Neuron-Attention decomposition. *arXiv preprint arXiv:2509.19943*, 2025.

Marc E Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 857–878, 2025.

Lawrence Chan, Adria Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In *AI Alignment Forum*, volume 2, pp. 19, 2022.

Gerasimos Chatzoudis, Konstantinos D Polyzos, Zhuowei Li, Difei Gu, Gemma E Moran, Hao Wang, and Dimitris N Metaxas. Can Cross-Layer transcoders replace vision transformer activations? an interpretable perspective on vision. *arXiv preprint arXiv:2604.13304*, 2026.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting Bi-Modal and Encoder-Decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 387–396. IEEE, 2021a.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791. IEEE, 2021b.

Haozhe Chen, Carl Vondrick, and Chengzhi Mao. Selfie: Self-Interpretation of large language model embeddings. *arXiv preprint arXiv:2403.10949*, 2024.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, volume 2024, pp. 2632–2652, 2024.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with Amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Parsa Esmaeilkhani and Longin Jan Latecki. Logit lens supervision for Patch-Level explanations in Vision-Language models. In *ECCV 2026 Workshop on Explainable Computer Vision: Challenges and Opportunities in the Era of Foundation Models (Proceedings Track)*, 2026.

Yossi Gandelsman, Alexei Efros, and Jacob Steinhardt. Interpreting clip's image representation via Text-Based decomposition. In *International Conference on Learning Representations*, volume 2024, pp. 18395–18416, 2024a.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting the Second-Order effects of neurons in CLIP. *arXiv preprint arXiv:2406.04341*, 2024b.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7457–7476, 2022.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer Feed-Forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, 2022.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, et al. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026.

Sangyu Han, Yearim Kim, and Nojun Kwak. Causal interpretation of sparse autoencoder features in vision. *arXiv preprint arXiv:2509.00749*, 2025.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *arXiv preprint arXiv:2403.17806*, 2024.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Lianghuan Huang and Yingshan Chang. Causality $\neq$ decodability, and vice versa: Lessons from interpreting counting ViTs. *arXiv preprint arXiv:2510.09794*, 2025.

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.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing Vision-Language representations to mitigate hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 63582–63605, 2025.

Sehyeong Jo, Gangjae Jang, and Haesol Park. GMAR:: Gradient-Driven Multi-Head attention rollout for vision transformer interpretability. In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 582–587. IEEE, 2025.

Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video. *arXiv preprint arXiv:2504.19475*, 2025.

Jinyeong Kim, Junhyeok Kim, Yumin Shim, Joohyeok Kim, Sunyoung Jung, and Seong Jae Hwang. Interpreting vision transformers via residual replacement model. *Advances in Neural Information Processing Systems*, 38:55760–55820, 2026.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020.

Alexander Lappe and Martin Giese. Register and [[CLS]] tokens induce a decoupling of local and global features in large ViTs. *Advances in Neural Information Processing Systems*, 38:1009–1029, 2026.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. *Advances in Neural Information Processing Systems*, 37:109233–109282, 2024.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, volume 2025, pp. 23888–23923, 2025.

Alexis Marouani, Oriane Siméoni, Hervé Jégou, Piotr Bojanowski, and Huy Vo. Revisiting [[CLS]] and patch token interaction in vision transformers. In *International Conference on Learning Representations*, volume 2026, pp. 22245–22263, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. *arXiv preprint arXiv:2407.08734*, 2024.

Mohammed Suhail B Nadaf. Steerable but not decodable: Function vectors operate beyond the logit lens. *arXiv preprint arXiv:2604.02608*, 2026.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in Vision-Language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

nostalgebraist. interpreting GPT:: the logit lens. LessWrong blog post, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2026-09-20.

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. *arXiv preprint arXiv:2304.07193*, 2023.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 548–560, 2023.

Alexander Pan, Lijie Chen, and Jacob Steinhardt. Latentqa: Teaching LLMs to decode activations into natural language. In *International Conference on Learning Representations*, volume 2026, pp. 78726–78759, 2026.

Felipe Parodi, Jordan K Matelsky, and Melanie Segado. Zero-Ablation overstates register content dependence in DINO vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4164–4170, 2026.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE:: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Anirudh Phukan, Harshit Kumar Morj, Apoorv Saxena, Koustava Goswami, et al. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in VLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9661–9675, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*, 2022.

Cody Rushing and Neel Nanda. Explorations of Self-Repair in language models. *arXiv preprint arXiv:2402.15390*, 2024.

Ronen Schaffer, Tomer Michaeli, and Meir Yossef Levi. From attention to prediction maps: Per-Class Gradient-Free transformer explanations. *OpenReview preprint*, 2025.

Aryan Sharma, Cutter Dawes, and Shivam Raval. Dissociating decodability and causal use in Bracket-Sequence transformers. *arXiv preprint arXiv:2604.22128*, 2026.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. *arXiv preprint arXiv:2602.22394*, 2026.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M Blei. Hypothesis testing the circuit Hypothesis in LLMs. *Advances in Neural Information Processing Systems*, 37:94539–94567, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

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. *arXiv preprint arXiv:2106.10270*, 2021.

Smitha Muthya Sudheendra and Jaideep Srivastava. When decodability is not enough: Logical validity representations, behavioral dissociation, and causal tests in language models. *arXiv preprint arXiv:2609.02438*, 2026.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024.

Ryota Takatsuki, Sonia Joseph, Ippei Fujisawa, and Ryota Kanai. Decoding vision transformers: The diffusion steering lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4819–4824. IEEE, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP:: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Martina G Vilas, Timothy Schaumlöffel, and Gemma Roig. Analyzing vision transformers for image classification in class embedding space. *Advances in Neural Information Processing Systems*, 36: 40030–40041, 2023.

Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7486–7496. IEEE, 2023.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM:: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 111–119. IEEE, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do Llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, 2024.

Junyi Wu, Bin Duan, Weitai Kang, Hao Tang, and Yan Yan. Token transformation matters: Towards faithful post-hoc explanation for vision transformer. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10926–10935. IEEE, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126 (10):1084–1102, 2018.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision*, pp. 696–712. Springer, 2022.

Nina Żukowska, Wolfgang Stammer, Bernt Schiele, and Jonas Fischer. Seeing through circuits: Faithful mechanistic interpretability for vision transformers. In *European Conference on Computer Vision*, pp. 164–181. Springer, 2026.

Appendix: Table of Contents

A Limitations **17**

B Extended related work **17**

C Reproducibility details **18**

D In-distribution and out-of-distribution CLS states **19**

E Path decomposition: how zero-ablation damage travels **20**

F Read window and content-light CLS **21**

F.1 Read-window criterion 21

F.2 Content-light CLS criteria 22

F.3 Scale probe on ViT-L 22

G RTLens details: operators, protocols, and robustness **23**

G.1 Complete message operator and first-order transport 23

G.2 Robustness and validity boundary 24

G.3 Executed-read and Jacobian-lens protocols 24

G.4 Per-patch labeling protocol 25

G.5 Logit-lens trajectories and readout dependence 26

H Attention map details: exact decomposition, design grid, and causal verification **27**

H.1 Exact decomposition and attribution scope 27

H.2 Design choices and read-window boundary 28

H.3 Causal verification and intervention dependence 29

I Full saliency benchmark **32**

I.1 Deletion faithfulness 33

I.2 Localization, concentration, and class specificity 34

I.3 External validity on ImageNet-S 35

A LIMITATIONS

This section discusses limitations that could not be included in the main text due to the page limit. We highlight three such limitations.

First, our causal anatomy depends on the choice of readout. This paper focuses on the standard ViT classification setting in which the classifier reads CLS rather than patch tokens directly. We therefore follow standard evaluation practice (Oquab et al., 2023; Siméoni et al., 2025) and use CLS-based readouts across all five ViT backbones. However, a different readout can change the relevant causal path. For example, an average-pooling probe or a dense prediction head reads patch tokens directly and may induce a different anatomy. Under a mean-pool readout on DINOv2, blocking $P \leftarrow P$ costs 73.8pp while blocking $C \leftarrow P$ costs only 0.1pp (§G), so the read window is specific to a CLS bottleneck and the construction does not transfer unchanged to pooled or attention-pooled readouts. We leave such readout-dependent anatomies to future work.

Second, the causal benchmark has limited resolution. The window-matched replacement test separates our map from the all-layer DLA map in three of five models, while the other two remain statistical ties. Input-level patch replacement does not consistently favor either map because it perturbs the entire downstream computation rather than isolating the measured read. More fine-grained causal comparisons between attribution maps will require higher-resolution interventions than accuracy drops from patch replacement.

Third, our map is not the best-performing method under every evaluation criterion. On standard saliency benchmarks that do not use our replacement intervention, such as deletion faithfulness, the gradient-based GAE often performs best. In addition, sparsity-matched controls indicate that part of our advantage in energy localization arises from the stronger spatial concentration of our map (§I).

B EXTENDED RELATED WORK

Anatomy of ViT classification. The closest prior locus result is Gandelsman et al. (2024a), who find that late attention layers dominate CLIP's image representation using mean ablation of direct effects. Related observational work reports non-parallel CLS processing (Lappe & Giese, 2026), late CLS aggregation through class-space projections (Vilas et al., 2023), and supervision-dependent attention patterns (Walmer et al., 2023), while register-token behavior (Darcet et al., 2024; Jiang et al., 2026) enters our path decomposition. We extend these observations with distribution-preserving interventions at token-group granularity, a path decomposition, and validation across non-CLIP backbones, native readouts, and ViT-L, yielding a causal account of the terminal read and a measured definition of the read window.

Intervention methodology. Our resampling interventions follow activation-patching guidance (Heimersheim & Nanda, 2024; Zhang & Nanda, 2024). This choice matters because ablation semantics can change circuit-level conclusions (Miller et al., 2024; Li & Janson, 2024), self-repair can mask component importance (Rushing & Nanda, 2024), and subspace interventions can activate dormant pathways and produce misleading localization (Makelov et al., 2024). We therefore favor on-distribution, operator-matched interventions and specify predictions before execution.

Lenses. The logit lens (nostalgebraist, 2020) reads intermediate language-model states through the frozen output head, while the tuned lens (Belrose et al., 2023) adds learned per-layer translators. Other variants use learned or model-based decoders, including Future Lens (Pal et al., 2023), Patchscopes (Ghandeharioun et al., 2024), SelfIE (Chen et al., 2024), LatentQA (Pan et al., 2026), and the Diffusion Steering Lens (Takatsuki et al., 2025), and several analyses rely directly on pre-read decodability (Wendler et al., 2024). The same idea has been applied to visual tokens in VLMs (Neo et al., 2025) and to lens-guided editing (Jiang et al., 2025), although naive logit-lens readouts can be unreliable for hallucination grounding (Phukan et al., 2025) and patch-level vocabulary alignment may need to be explicitly trained (Esmaeilkhani & Latecki, 2026). In ViTs, intermediate CLS states have been projected into class space (Vilas et al., 2023), and vision interpretability toolkits expose the same readout (Joseph et al., 2025). Yet ViTs route patch content to the classifier through late attention reads, and lens designs that account for this causal path remain underexplored. This concern is consistent with prior causal work showing that decodable features need not be behaviorally used (Elazar et al., 2021; Canby et al., 2025; Han et al., 2025).

DLA and circuit lineage. Prior work decomposes attention outputs (Kobayashi et al., 2020) and interprets representations in CLIP (Gandelsman et al., 2024a), non-CLIP ViTs (Balasubramanian et al., 2024), and ResNet-based CLIP (Bu & Gandelsman, 2025). We extend this line with classifier-space attribution for non-CLIP models, exact per-image bookkeeping with unattributed terms, and semantics-matched intervention. Beyond attention decomposition, circuit analysis has progressed toward sparse feature circuits (Marks et al., 2025), transcoders (Dunefsky et al., 2024), cross-layer attribution graphs (Ameisen et al., 2025; Lindsey et al., 2025), approximate edge attribution (Syed et al., 2024; Hanna et al., 2024), and formal tests of circuit hypotheses (Shi et al., 2024). Vision work has begun to port these ideas through edge-level circuit discovery (Żukowska et al., 2026), cross-layer transcoders (Chatzoudis et al., 2026), and residual replacement models (Kim et al., 2026). Our analysis instead operates at the coarser level of token-group edges, where causal effects can be measured directly.

Concurrent work. Two concurrent lines overlap with parts of our negative result. Huang & Chang (2025) report a decodability-use dissociation in counting ViTs, and Parodi et al. (2026) find register and sink behavior consistent with parts of our path decomposition. Our contribution differs in establishing the terminal read and content-light placeholder through operator-controlled interventions, reproducing this anatomy across five backbones and native readouts, and deriving instruments from measured validity windows. Further concurrent reports of lazy CLS aggregation (Shi et al., 2026), token-specific LayerNorm structure (Marouani et al., 2026), and concept-specialized attention heads (Basile et al., 2026) are consistent with the division of labor we identify. Related decodability-use dissociations have also been reported concurrently in language models and controlled transformers (Nadaf, 2026; Sudheendra & Srivastava, 2026; Sharma et al., 2026).

C Reproducibility details

Evaluation scale. Table 4 summarizes the sample size of each experiment. For the single-edge screen with 4,000 images, cells below 1pp in Table 1 have binomial half-widths of at most 0.31pp.

Table 4: Sample sizes by experiment.

Experiment	Section	Sample
Linear probe training	§2	20,000 images
Joint pre-read CLS replacement and lens trajectories	§3.2	5,000 images
Nine-edge screen (single-layer and suffix cuts)	§3.1	4,000 images
Dataset-mean CLS replacement	§3.2	2,000 images
Native readouts and DINOv3 pre-read CLS replacement	§3.2	2,000 images
Path decomposition and ViT-L	§E, F.3	2,000 images
Margin decomposition and head profiles	§H	1,000 images
Saliency benchmark	§5.2, I	500 images
Pointing subsets (box and correctness filtered)	§I	362–421 images
Jacobian lens estimation	§4.2	128 images
Per-neuron clamp test (patch/CLS)	§4.2	48/64 neurons per layer, 256 images

Deterministic data splits. All ImageNet-1k splits are deterministic. We seed a shuffle of the validation set, use the first 20,000 images for probe training, and draw every evaluation subset from the remaining images. Probe-training and evaluation images are therefore disjoint.

Probe construction. Unless stated otherwise, the readout is a linear head trained on frozen final-norm CLS features. The probe is optimized with Adam for 300 full-batch steps using a learning rate of 10^{-3}, weight decay 10^{-4}, standardized inputs, and a fixed seed. Each experiment retrains the probe from this fixed recipe within its own run rather than loading a shared checkpoint, ensuring that the readout is matched to that run's extracted features. Since both the evaluation subset and the resulting probe instance can differ across runs, clean accuracy varies across experiments. All intervention effects are therefore reported relative to the clean baseline from the same run.

Native and experiment-specific readouts. The supervised ViT experiments use the timm checkpoint `augreg2_in21k_ft_in1k`. The CLIP zero-shot readout uses an 80-template prompt ensemble and achieves clean top-1 accuracy 0.679. Following the readout policy of §2, the causal interventions for CLIP and the supervised models, including their rows in Table 1, are scored with their native

readouts on the same 4,000-image evaluation subset as the screen, with the supervised readout path asserted to reproduce the models' own forward logits. For the DINOv2 pre-read CLS intervention series, we additionally train a separate CLS head with AdamW for five epochs on 100,000 images using random crops and horizontal flips. The linear probe reproduces the headline pre-read CLS dissociation, with a resampling effect of 4.7pp on 2,000 images.

Donor construction. Replacement interventions replace part of a target image's internal representation with the corresponding content from another image, which we call the *donor*. Each target is paired with a donor from the same batch. Edge interventions replace the post-softmax contribution along the targeted edge with the donor's corresponding contribution, selected by a seeded within-batch permutation. State and path interventions use a one-position batch roll, pairing each image with the next image in the batch. Saliency deletion averages over two donor assignments obtained from batch rolls of 1 and 7, while the imputation, insertion, and randomization evaluations use only the roll-1 donor.

Bootstrap granularity. Saliency results (§5.2, I) are averaged over the two donor assignments described above, whereas the remaining intervention experiments use a single seeded donor assignment. For saliency deletion, donor choice directly affects which content replaces the selected patches, and with only 500 images an atypical pairing could noticeably affect comparisons between methods. We therefore average over two donor assignments. The remaining interventions are evaluated over thousands of images and produce effects that are large relative to donor variation, so a single seeded assignment is sufficient. Bootstrap confidence intervals preserve these donor assignments and therefore capture variation over the evaluated images, not over an additional distribution of possible donor pairings.

Probe-seed variation. We separately sweep three probe seeds for the headline DINOv2 pre-read CLS dissociation (§3.2). With 2,000 images, clean accuracy is 79.2±0.1%, the pre-read CLS resampling effect is 5.9±0.2pp, and the zero-ablation effect is 74.7±0.8pp. Variation across evaluation subsets is larger than variation across probe seeds, indicating that the headline dissociation is stable to the tested probe seeds.

Experiment coverage. To preclude selective-reporting ambiguity and to make differences in coverage explicit, we record in Table 5 which models each main-text experiment was run on. A ✓ denotes an experiment that was run and reported, — denotes one that was not run, n/a denotes an experiment that is structurally inapplicable, and an edge label denotes coverage restricted to that edge. Coverage for appendix-only analyses is stated in the corresponding appendix sections.

Table 5: Experiment coverage across models.

Experiment	Section	DINOv2	DINOv3	CLIP	sup. ViT	DeiT
Nine-edge screen + frozen singles	§3.1	✓	✓	✓	✓	$C \leftarrow P$
Pre-read CLS state interventions (resample/zero)	§3.2	✓	✓	✓	✓	✓
Native-readout verification	§3.2	n/a	n/a	✓	✓	✓
Logit-lens trajectory	§3.2	✓	✓	✓	✓	✓
Tuned lens	§3.2	✓	✓	✓	✓	✓
CLS-position clamp test	§4.2	✓	✓	✓	✓	✓
Patch-position clamp test (read-through)	§4.2	✓	✓	✓	✓	✓
Exact decomposition + patch replacement	§5.1	✓	✓	✓	✓	✓

D IN-DISTRIBUTION AND OUT-OF-DISTRIBUTION CLS STATES

The interventions of §3.2 replace pre-read CLS states with three kinds of substitutes. Cross-image resampling stays on the model's own activation distribution, the dataset-mean state preserves its channel-wise scale while carrying no image content, and zeroing produces a state the backbone never emits. On DINOv2, the first two substitutions cause only small drops, while zeroing causes a 69.1pp drop (Fig. 4a). The zeroing effect also varies sharply across models and falls to 19.8pp on DINOv3, which is difficult to explain if zeroing measured the removal of image-specific content.

To separate distributional validity from content, we construct a state that carries no image information but respects the channel statistics. We estimate the mean and standard deviation of each CLS channel

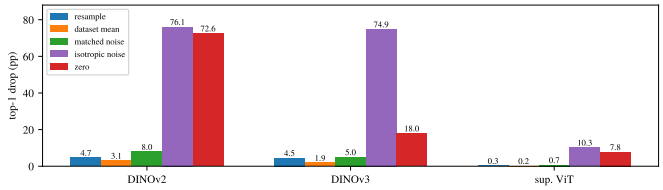


Figure 9: In-distribution and out-of-distribution CLS replacements below the read window (n=2,000, all five arms from the same run per model; L0–10 for DINOv2 and DINOv3 with the linear probe, L0–9 for supervised ViT with its native head). Replacements that respect the channel-wise statistics (resampling, dataset mean, matched noise) cause small drops, while replacements that leave the activation distribution (isotropic noise, zeros) are destructive. The zeroing effect differs sharply across models, whereas the matched versus isotropic contrast is shared.

and sample Gaussian noise matched to these channel-wise statistics. The results are shown in Fig. 9. On DINOv2, this matched noise causes only an 8.0pp drop, comparable to cross-image resampling. Isotropic noise with the same overall scale, which destroys the channel-wise profile, is instead as destructive as zeroing and produces a 76.1pp drop. The same contrast holds on DINOv3 despite its much smaller zeroing effect. On supervised ViT, matched noise costs only 0.7pp while isotropic noise costs 10.3pp, exceeding zeroing (7.8pp). Before the read window, the model therefore requires the CLS position to hold *an in-distribution activation profile rather than image-specific content*.

E PATH DECOMPOSITION: HOW ZERO-ABLATION DAMAGE TRAVELS

§D shows that the large effect of zeroing arises from distributional disruption. Here we ask how this damage travels through the model. A token can affect attention through two communication channels. As a *query*, it determines what the token reads from other tokens. As a *key/value*, it determines what other tokens read from it. Its state can also persist through the residual stream without passing through attention. To distinguish these paths, we independently zero or resample the query channel, the key/value channel, and the full token state below the read window of DINOv2 (*i.e.*, L0–10).

The interventions act on the projected per-head vectors inside attention, after query and key normalization when present. Writing q_t^h, k_t^h, and v_t^h for the query, key, and value of token t in head h, the query intervention replaces q_t^h for tokens in the target group, while the key/value intervention replaces both k_t^h and v_t^h for the same tokens. The state intervention instead replaces the block output at the corresponding token position. Zeroing sets the selected vectors to zero, while resampling substitutes the corresponding vectors from the donor image. Attention weights are recomputed using the modified queries or keys for query and key/value interventions, respectively.

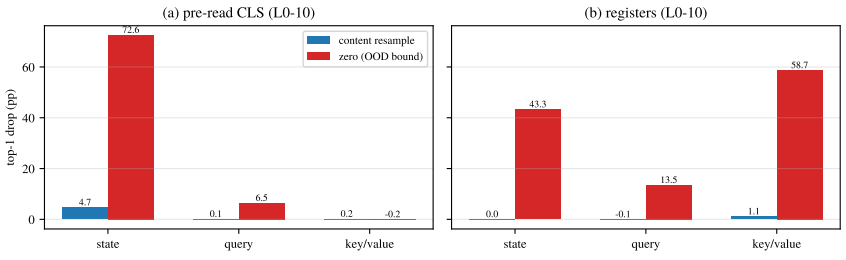


Figure 10: Path decomposition of pre-read token interventions on DINOv2 over L0–10 (n=2,000). **(a)** For CLS, neither attention channel reproduces the large damage caused by zeroing the full state, while content resampling is harmless in every channel. **(b)** For registers, the key/value channel carries nearly the full zero-ablation effect, while cross-image resampling of the same content remains harmless.

CLS damage travels through residual continuity, not communication. Fig. 10a shows that zeroing either the CLS query channel or its key/value channel causes only a small accuracy drop compared with zeroing the full CLS state. Cross-image resampling is likewise largely harmless across the state,

Table 6: Register-removal operators on DINOv2 over L0–10. Attention share is the fraction of attention mass received by register tokens (n=512 images); drops are top-1 accuracy drops (n=2,000). Value-only operators leave the attention distribution unchanged. The paired gap between softmax removal and key/value zeroing is +4.5pp [+3.5, +5.5].

Register operator	att. share (%)	top-1 drop (pp)
none (clean)	23.8	–
full-state zero	13.6	43.3
key/value zero	1.5	58.7
softmax removal	0.0	63.2 [61.1, 65.4]
value-only zero	23.8 (unchanged)	75.8 [73.9, 77.7]
value-only resample	23.8 (unchanged)	0.4 [-0.3, 1.1]

query, and key/value channels. The large zero-ablation effect therefore does not propagate through CLS attention. Instead, the disrupted CLS state persists through the residual stream until the terminal read and final normalization. These results show that *pre-read CLS is required as an in-distribution residual state*, but not as an important communication node.

Registers matter as attention targets, not through image-specific content. Registers show a different routing pattern. As shown in Fig. 10b, zeroing their key/value channel accounts for most of the damage and is even more disruptive than zeroing the full state, whereas replacing the same key/value content across images has little effect. To test whether this effect is specific to the key/value-zeroing operator, we compare alternative register-removal interventions in Table 6. Zeroing register keys nearly eliminates their attention share, and removing registers from the softmax entirely is more damaging still, showing that key/value zeroing does not inflate the effect. Value-only zeroing is even more destructive because registers continue to absorb attention while returning no content, whereas value resampling is largely harmless. These results show that *registers are required primarily as attention targets*, while their image-specific content is largely interchangeable.

F Read window and content-light CLS

The main text identifies a late *read window*, where patch information is transferred into CLS and becomes available to the classifier. This appendix formalizes that organization and tests how far it generalizes. We first define the read-window criterion that separates single-read from dual-read models (§F.1). We then formalize the content-light behavior of pre-read CLS states and show why the measured read boundary is essential to evaluating it (§F.2). Finally, we probe whether the same organization persists at larger scale on DINOv2 ViT-L/14 (§F.3).

F.1 Read-window criterion

We distinguish models whose read is concentrated in the final layer from those whose read extends across the final two layers. Let $\mathrm{effect}(\cdot)$ denote the top-1 accuracy drop caused by an intervention relative to its clean baseline. We compare the joint intervention on L10–11 with the stronger of the two single-layer interventions and define the *joint-minus-single gap* as

$$\mathrm{gap} = \mathrm{effect}(\text{L10–11}) - \max\big(\mathrm{effect}(\text{L10}), \mathrm{effect}(\text{L11})\big). \tag{6}$$

We include L10 in the read window when this gap exceeds half of the strongest single-layer effect:

$$\mathrm{gap} > \frac{1}{2}\max\big(\mathrm{effect}(\text{L10}), \mathrm{effect}(\text{L11})\big). \tag{7}$$

This criterion cleanly separates the two regimes. For DINOv2, DINOv3, and CLIP, the gaps are 0.1, 0.6, and 2.7pp, respectively, each at most 0.15 times the strongest single-layer effect. We therefore define their read window as L11. In contrast, supervised ViT and DeiT have gaps more than 2.5 times their strongest single-layer effects, indicating that both L10 and L11 contribute substantially to the read. Their read window is therefore L10–11. The two regimes are well separated, making the classification insensitive to the exact cutoff.

The absolute intervention effects depend on how the read is perturbed. For example, on supervised ViT, edge resampling and fixed-weight value swapping produce L11 drops of 17.0 and 33.0pp,

whereas on DINOv2 the corresponding effects are 78.9 and 76.5pp. This difference is expected because edge resampling replaces the donor attention contribution, while fixed-weight value swapping preserves the target attention weights and replaces only the value content. We therefore compare single-layer and multi-layer effects only within the same intervention scheme. Under fixed-weight value swapping, which matches the read decomposition in §5.1, DeiT again exhibits a clear two-layer read, with an L11 effect of 17.5pp compared with 61.3pp for L10–11. The read-window distinction therefore does not depend on edge resampling alone.

F.2 Content-light CLS criteria

By *content-light*, we do not mean that pre-read CLS states contain no class information. These states can remain strongly decodable, as shown in Fig. 4b,c and Fig. 14. The claim is instead that downstream computation makes little use of their image-specific content before the read window. We define this behavior using two comparisons. First, replacing pre-read CLS content with content from another image should be no more damaging than replacing it with a statistics-matched random state. Second, content replacement should be substantially less damaging than destroying the CLS state itself. Excluding the read window leaves L0–10 for single-read models and L0–9 for dual-read models. Across these ranges, content replacement causes only a small accuracy drop, while zeroing the same states is substantially more damaging (Fig. 4a). Before the read window, the model therefore requires a plausible CLS state while depending much less on the particular image-specific content carried by that state. We regard the content-light criterion as violated when cross-image replacement is more damaging than the matched-random control, regardless of the absolute effect size.

The measured read boundary is essential to this evaluation. In the dual-read models, whose read window is L10–11, replacing CLS content over L0–9 has essentially no effect (Fig. 4a). Extending the same intervention by a single layer to L0–10 increases the effect to 3.5pp on supervised ViT and 13.9pp on DeiT. The intervention now overlaps the first read layer, where CLS content is already being causally used. *A fixed layer range chosen without reference to the measured read window would therefore misclassify both models as violating the content-light criterion.* CLIP illustrates the complementary case. Its read remains concentrated at L11 under the criterion above. Although L10 makes a modest secondary contribution, the joint-minus-single gap remains small relative to the effect of intervening on L11 alone. Its pre-read range is therefore still L0–10.

F.3 Scale probe on ViT-L

We next ask whether the same organization persists at larger scale using the 24-layer DINOv2 ViT-L/14. The corresponding results are shown in Fig. 11. The overall organization remains intact, but the read becomes wider. Patch-to-CLS effects remain concentrated in the final layers, with L21–23 accounting for 93.1% of the summed positive $C \leftarrow P$ effect, while register-to-CLS cuts remain negligible. The logit-lens trajectory and clamp-test behavior also change over this same late region, providing convergent evidence that the read extends across a broader terminal range.

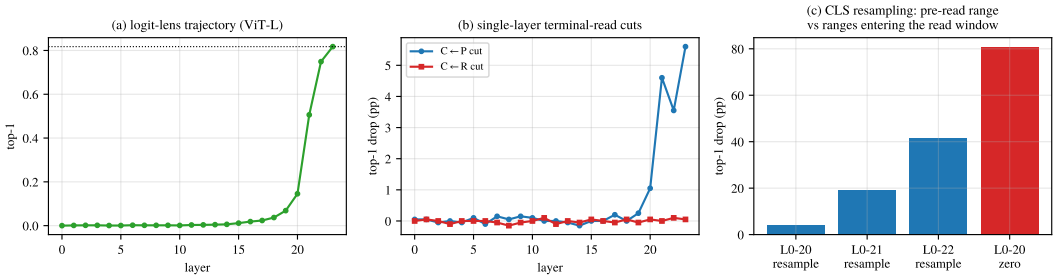


Figure 11: Scale probe on DINOv2 ViT-L/14. **(a)** The logit-lens trajectory changes over L21–23. **(b)** Terminal $C \leftarrow P$ effects concentrate in the final layers, while register-to-CLS effects remain near zero. **(c)** Replacing pre-read CLS content is nearly harmless (4.3pp), whereas extending the intervention into the read window is strongly damaging. The placeholder pattern therefore persists at scale with a wider read window.

The content-light pattern persists as well. Replacing CLS content over the pre-read range L0–20 causes only a 4.3pp drop, while extending the same intervention into the late read region is substantially more damaging. Therefore, *the primary scale-dependent change is the width of the read: ViT-B concentrates the read in the final one or two layers, while ViT-L spreads it across roughly the final three or four layers.* In both cases, patch representations develop before the read and become causally relevant to the classifier once their information is transferred into CLS. This provides an initial causal test of the scale-dependent global-token behavior described observationally by Lappe & Giese (2026).

Taken together, the read-window and content-light criteria capture complementary aspects of the same organization. The former identifies when patch information becomes causally available to the classifier, while the latter asks whether image-specific CLS content matters before that transfer. Across architectures and model scales, pre-read CLS states can remain decodable while carrying little causally used image-specific content. The main variation lies in how many terminal layers participate in the read.

G RTLens details: operators, protocols, and robustness

The main text defines RTLens by transporting intermediate patch states through the read window before applying the classifier. This appendix gives the full operators, characterizes when the transport is causally reliable, and specifies the analysis protocols. We first restore the complete message operator and define the first-order transport used in the clamp test (§G.1). We then map the validity boundary of that approximation across LayerNorm terms, clamp strength, and write depth (§G.2). We next describe the executed-read trajectory and ported Jacobian-lens protocols (§G.3) and the per-patch labeling protocol (§G.4). Finally, we clarify what the naive lens trajectories measure and how their reading depends on the readout (§G.5).

G.1 Complete message operator and first-order transport

The main text omits bias and LayerScale terms from the message sum in Eq. (2). Restoring them gives

$$\sum_{r\in\mathcal{W}_\ell}\Big(\textstyle\sum_n \mathrm{msg}^r(P_n^{(\ell)}) + \mathrm{ls}_1^r \odot b_O^r\Big). \tag{8}$$

For a read layer r evaluated on source states $P^{(\ell)}$, the complete message operator shared by RTLens and the map s_n of §5.1 is

$$\mathrm{msg}^r\big(P_n^{(\ell)}\big) \coloneqq \mathrm{ls}_1^r \odot \textstyle\sum_h A^{r,h}[\mathrm{CLS},n]\, W_O^{r,h}\, v_n^{r,h}, \qquad v_n^{r,h} = W_V^{r,h}\,\mathrm{LN}_1^r\big(P_n^{(\ell)}\big) + b_V^{r,h}. \tag{9}$$

Here, ls_1^r is the LayerScale gain, equal to the identity when absent, and $b_V^{r,h}$ is the value bias. When the source-layer superscript is omitted, the message is evaluated on the actual input to layer r. The constant output bias $\mathrm{ls}_1^r \odot b_O^r$ is added once for each executed read layer. Since these terms enter additively through the residual stream, the same operator supports both the window sum in Eq. (2) and the window-restricted attribution map.

The direction-level transport of Eq. (4), used in the clamp test of §4.2, *makes a first-order approximation to this state-level operator*. It removes the realized per-head attention mass and per-token normalization scale from Eq. (9) and retains only the linear transport of the write direction, yielding

$$T_r w = \mathrm{ls}_1^r \odot \textstyle\sum_h W_O^{r,h} W_V^{r,h}\big(\gamma_1^r \odot (w - \bar{w}\,\mathbf{1})\big), \tag{10}$$

where γ_1^r is the pre-attention LayerNorm gain and $\bar{w}\,\mathbf{1}$ performs centering. In this notation, the neuron lens of Eq. (4) reads

$$\mathrm{RTLens}(w) = \arg\max_c z_c\Big(\textstyle\sum_{r\in\mathcal{W}_\ell} T_r w\Big). \tag{11}$$

The clamp test sums T_r over the remaining read layers in $\mathcal{W}_\ell$. For single-read models, this reduces to the terminal operator. For dual-read models, both read operators are included below L10. The state-level RTLens of Eqs. (2) and (3) does not make this direction-level approximation.

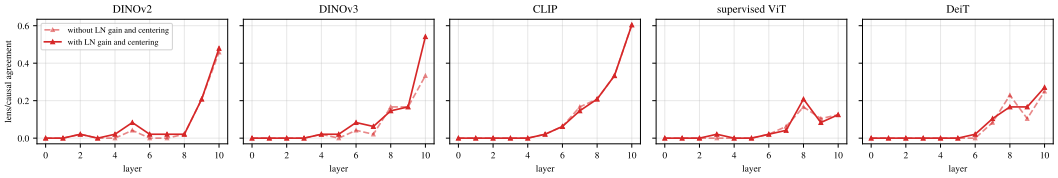


Figure 12: LayerNorm-diagonal ablation of the direction transport. Clamp agreement under the terminal transport with the LayerNorm gain and centering included (solid) or removed (dashed), across the five backbones.

G.2 Robustness and validity boundary

The direction-level transport is a first-order approximation, so we test when it remains causally reliable by varying the terms retained in the approximation, the strength of the clamp intervention, and the depth of the transported write. The LayerNorm terms in Eq. (10) matter primarily on DINOv3. At L10, including the gain and centering raises agreement from 0.33 to 0.54, while the effect is negligible on the other backbones. Fig. 12 shows both variants.

Dual-read models introduce an additional source of approximation error because a write below the read window must pass through more than one read layer. Summing over the full read window partially recovers this loss. At L9, agreement rises from 0.08 to 0.15 on supervised ViT and from 0.17 to 0.27 on DeiT. Across all pre-read layers, paired comparisons favor the full-window sum on 23 neurons against 6 under the same interventions, with a two-sided sign test of $p{=}0.002$. The remaining decline at earlier layers therefore cannot be explained by omitting the second read layer.

The read-through advantage is also stable across clamp strengths. Under q95, q99, and q99.9, L10 agreement is 0.46/0.48/0.71 on DINOv2 and 0.56/0.60/0.54 on CLIP, remaining well above the null at every dose. The 1,000-way argmax criterion is conservative as well. At L10, the causally promoted class appears within the read-through label's top 5 for 0.79/0.85/0.98 of neurons on DINOv2/DINOv3/CLIP. The lower agreement of dual-read models is not explained by sample size. With $n{=}128$ neurons, DeiT reaches 0.30 [0.22, 0.38] and supervised ViT reaches 0.09 [0.05, 0.15], both below the single-read models. Together, these results indicate that the first-order read-through transport is *most reliable when the causal read is concentrated in a single terminal layer*.

Agreement nevertheless declines at earlier depths, where neuron writes undergo additional $P \leftarrow P$ mixing before reaching the read window. We therefore test whether adding one linearized $P \leftarrow P$ hop extends the range over which the transport remains causally predictive. It does not. At L9, two-hop agreement falls from 0.21 to 0.06 on DINOv2 and from 0.17 to 0.02 on DINOv3, while CLIP is essentially unchanged at 0.31 versus 0.33. No model gains a wider validity range from the additional hop. These results place the practical validity boundary of the first-order approximation near the read window. Once additional patch-to-patch computation intervenes, a single linearized hop is insufficient to capture the downstream transformation. This explains why the read-through transport is causally reliable near the read window but not uniformly across depth.

G.3 Executed-read and Jacobian-lens protocols

The trajectory in Fig. 6 uses the state-level RTLens and directly applies the read window to layer-ℓ patch states through Eq. (2). Each read layer uses its realized pre-attention LayerNorm statistics, value projections and biases, CLS attention row, output projection, and LayerScale. In dual-read models, both remaining read operators are evaluated when both lie above layer ℓ, while only the terminal read applies at $\ell{=}10$. The corresponding fidelity ablation replaces this state-level execution with the first-order operator of Eq. (10), which omits the realized per-token LayerNorm scale and attention mass. The first-order operator substantially reduces accuracy, particularly on CLIP and the dual-read models. At L10, for example, CLIP reaches 0.51 under state-level execution versus 0.08 under the first-order approximation. A high-norm control that excludes patches above $2\times$ the per-image median norm changes CLIP accuracy by at most 0.03 at L9–10. The improvement from state-level execution therefore comes from *retaining the realized normalization and attention*



Figure 13: Example per-patch label maps on CLIP. Green marks patches whose predicted label matches the image class. The naive logit lens labels almost no patches correctly, while the read-through projection recovers the object region.

weighting rather than from suppressing unusually large patch states. All trajectory variants use the same probe, split, and n=2,000 evaluation images as the logit-lens baseline.

As an independent comparison, we port the Jacobian lens of Gurnee et al. (2026). At each layer, we compute $J_\ell = \mathbb{E}[\partial h^\star / \partial h_\ell]$, where $h^\star$ is the final pre-norm CLS state and h_ℓ is the full layer state. We estimate the expectation over 128 images using full backpropagation, with no components treated as constants. Separate operators are computed for CLS and patch source positions, with patch Jacobians averaged over positions and images as in Gurnee et al. (2026). The frozen classifier is then applied to $J_\ell h$. Two implementation checks verify the construction. First, J_{11} is exactly the identity. Second, split-half matrix cosines are at least 0.96 at L9 and exceed 0.97 at L10 in all three ported models. These checks show that the averaged operators are numerically stable enough for the comparison reported in the main text.

G.4 Per-patch labeling protocol

The labeling experiment of §4.3 uses 500 evaluation images per model. For patch n at layer ℓ, the read-through label is

$$\arg\max_c \left\langle q_c, \sum_{r \in \mathcal{W}_\ell} \mathrm{msg}^r(P_n^{(\ell)}) \right\rangle. \tag{12}$$

Here, q_c is the residual-space class direction defined in §5.1. The projection uses the realized CLS attention rows and the same intercept-free class matrix as the clamp test, matching the per-patch form of Eq. (3). The naive logit-lens label instead applies the frozen readout directly to the patch state. Both methods therefore use the same intermediate representation and classifier weights, differing only in whether the measured read path is applied before classification.

Fig. 13 shows representative CLIP label maps. Table 7 gives L10 patch-stream neurons for which the read-through label matches the class promoted by clamping while the naive logit-lens label does not. These examples provide a unit-level view of the aggregate agreement difference in Fig. 5.

Table 7: Example L10 patch-stream MLP neurons from the clamp-test sample. For each neuron, the read-through label matches the causally promoted class, while the naive logit-lens label does not. Across the full L10 samples, read-through labels match the causal class for 0.60 of CLIP neurons and 0.54 of DINOv3 neurons, whereas the naive logit lens matches none of the sampled L10 neurons in these two models.

Model	Neuron (L10)	Causally promoted class = read-through label	Naive label
CLIP	86	Shih Tzu	sliding door
	977	mitten	binoculars
	751	ice cream	hammerhead shark
	706	abacus	bassinet
DINOv3	2232	frying pan	tray
	1969	guenon	coyote
	706	fox squirrel	water jug
	694	velvet fabric	hair clip

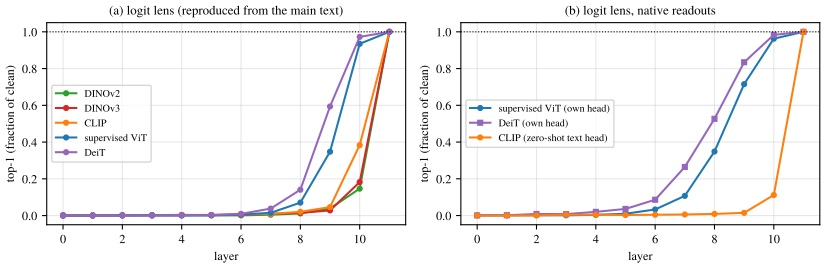


Figure 14: Logit-lens trajectories under **(a)** the linear probe, reproduced from Fig. 4c for reference, and **(b)** native readouts. Self-supervised models and CLIP rise sharply near the read window, while the supervised models rise earlier. Native readouts preserve the same objective-dependent pattern.

G.5 Logit-lens trajectories and readout dependence

Finally, we step back from the read-through construction and clarify what the naive lens trajectories themselves measure, since RTLens is motivated by and compared against them. Projecting intermediate CLS states through the final head (Vilas et al., 2023) produces markedly different trajectories across training objectives, as shown in Fig. 4c and reproduced for reference in Fig. 14a. The self-supervised models and CLIP remain near zero through L9 and rise sharply near the terminal read, with the self-supervised models showing the largest increase from L10 to L11. The supervised models rise earlier, and this qualitative pattern persists under their native readouts (Fig. 14b).

These trajectories nevertheless do not identify where the prediction is causally computed. In the self-supervised models, the sharp late rise coincides with the terminal $C \leftarrow P$ read and the loss of CLS content invariance, indicating the onset of alignment between patch evidence and the CLS readout. In the supervised models, by contrast, decodability emerges before causal use. DeiT is already strongly decodable at L9 even though replacing its pre-read CLS content has approximately no causal effect. A tuned lens (Belrose et al., 2023) rises earlier still because it learns a layer-specific translation. *Readout alignment, recoverability, and causal use are therefore distinct quantities, and only causal intervention establishes the last.*

The interpretation of a lens also depends on the readout through which the representation is evaluated. On supervised ViT, the native head rises one layer earlier than the linear probe, with substantial accuracy by L8. More importantly, changing the readout can alter the causal path itself. Under a mean-pool readout on DINOv2, blocking $P \leftarrow P$ reduces accuracy by 73.8pp, while blocking $C \leftarrow P$ reduces it by only 0.1pp. Intermediate projections should therefore be interpreted for a specific model–readout pair, since both the apparent lens trajectory and the relevant causal anatomy can change with the readout.

Taken together, these analyses separate two versions of read-through interpretation. The state-level RTLens uses the realized read-layer normalization and attention weights and can recover information

that depends on these image-specific quantities. The direction-level transport instead provides a simpler first-order approximation whose causal validity is concentrated near the read window. More broadly, decodability alone does not determine how an intermediate projection should be interpreted. Its meaning depends on the path through which the model consumes the representation, the depth over which the projection remains causally valid, and the readout against which it is evaluated.

H ATTENTION MAP DETAILS: EXACT DECOMPOSITION, DESIGN GRID, AND CAUSAL VERIFICATION

The main text derives a spatial attribution map by restricting direct logit attribution to the read window and scoring the patch content delivered along the classifier's decision direction. This appendix gives the exact decomposition behind that map and tests the design choices and causal validity of the resulting attribution. We first derive the exact additive decomposition and clarify which component of the prediction margin the map attributes (§H.1). We then examine the anatomy-derived design choices, the layer scope and the scored quantity, through a full design grid and a read-window boundary test (§H.2). Finally, we verify that high-scoring patches are causally important and show how this verification itself depends on the intervention operator (§H.3).

H.1 EXACT DECOMPOSITION AND ATTRIBUTION SCOPE

Let $h^\star$ denote the CLS state entering the final normalization, and let h_{pre} denote the CLS state entering the first layer of the read window $\mathcal{W}$. Because transformer blocks are residual, the final CLS state decomposes exactly into the state entering the window and the attention and MLP writes produced within it:

$$h^\star = h_{\mathrm{pre}} + \sum_{r\in\mathcal{W}} \Big(\sum_n \mathrm{msg}^r(P_n) + \mathrm{ls}_1^r \odot b_O^r + \mathrm{mlp}^r \Big), \tag{13}$$

The token sum runs over patches, CLS, and registers. For single-read models, $\mathcal{W}$ contains one block. For dual-read models, the identity unrolls both read layers.

Conditioned on the realized per-image normalization scale σ_I, the frozen readout is affine. The margin between the predicted class $\hat{c}$ and the runner-up c_2 can therefore be written as

$$\mathrm{margin} = z_{\hat{c}}(h^\star) - z_{c_2}(h^\star) = \langle h^\star, \tilde{q} \rangle + c(\hat{c}, c_2), \qquad \tilde{q} = q - \bar{q}\,\mathbf{1}, \quad q = \frac{\gamma}{s\,\sigma_I} \odot \big(W_{\hat{c}} - W_{c_2}\big). \tag{14}$$

The gain γ and shift β come from the final normalization, W_c and b_c are class c's readout row and bias, and s and μ_F are the per-channel scale and mean of the readout's input standardization. Mean-centering q absorbs the centering term of the final normalization exactly, because μ_I is the mean of $h^\star$ and $\tilde{q} \perp \mathbf{1}$. The intercept

$$c(\hat{c}, c_2) = \big\langle (W_{\hat{c}} - W_{c_2}) \oslash s,\ \beta - \mu_F \big\rangle + b_{\hat{c}} - b_{c_2} \tag{15}$$

collects the state-independent bias terms.

Substituting Eq. (13) into Eq. (14) gives the exact additive decomposition

$$\mathrm{margin} = \underbrace{\textstyle\sum_{n\in\mathrm{P}} s_n}_{\text{per-patch map}} + \underbrace{\textstyle\sum_{n\notin\mathrm{P}} s_n}_{\text{global-token messages}} + \langle h_{\mathrm{pre}}, \tilde{q} \rangle + \textstyle\sum_{r\in\mathcal{W}} \big(\langle \mathrm{mlp}^r, \tilde{q} \rangle + \langle \mathrm{ls}_1^r \odot b_O^r, \tilde{q} \rangle \big) + c(\hat{c}, c_2), \tag{16}$$

where s_n is the per-patch score used in §5.1. Grouping terms recovers the five-part decomposition from the main text, with $s_{\mathrm{glob}} = \sum_{n\notin\mathrm{P}} s_n$, $s_{\mathrm{res}} = \langle h_{\mathrm{pre}}, \tilde{q} \rangle$, $s_{\mathrm{MLP}} = \sum_{r\in\mathcal{W}} \langle \mathrm{mlp}^r, \tilde{q} \rangle$, and $s_{\mathrm{const}} = \sum_{r\in\mathcal{W}} \langle \mathrm{ls}_1^r \odot b_O^r, \tilde{q} \rangle + c(\hat{c}, c_2)$. *The identity holds per image up to numerical residuals below* 10^{-6} *in every model*, including the window-unrolled dual-read models.

This decomposition is an accounting identity for the realized forward pass, not a counterfactual claim. Perturbing one term therefore need not change the margin by exactly its attributed share, which is why we verify the map separately by intervention in §5.3. The spatial map attributes only the patch-delivered component of the margin, while the residual, global-token, MLP, and constant terms remain explicit in Eq. (16). For reference, Fig. 15 reports the per-image share of each component, showing how much of the margin is covered by the spatial map.

In dual-read models, the layer-11 term in s_n is evaluated on $P_n^{(10)}$, after layer 10 has already mixed information across patches. The state at slot n may therefore contain content gathered from other

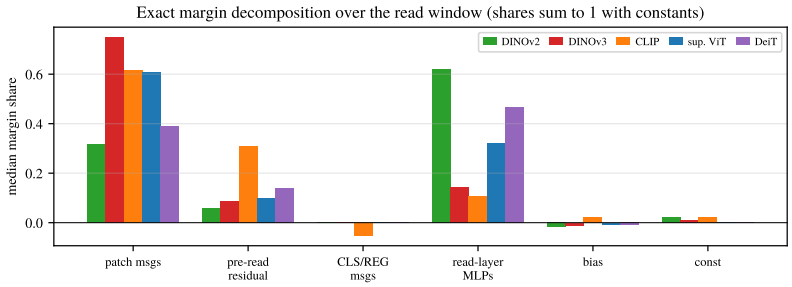


Figure 15: Per-model margin shares of the exact read-window decomposition in Eq. (16). The figure reports per-share medians, which need not sum to 1 even though the shares sum to 1 for each image. For dual-read models, patch and MLP terms include contributions from both read layers.

spatial locations. The map should thus be interpreted as attributing the margin to the token slot through which information is delivered at the read window, not necessarily to the corresponding input patch alone. The additive decomposition remains exact under this interpretation. This distinction also explains the intervention comparison in §5.3. Read-window value replacement removes the attributed delivery directly, whereas input-level replacement perturbs the patch before within-window mixing, so the two interventions need not agree. The DLA map should therefore be interpreted as an attribution of the attention read rather than of the entire classification decision.

H.2 Design choices and read-window boundary

The proposed map involves three distinct choices: the layer scope over which attribution is accumulated, the quantity used to score each attention message, and the class-space direction onto which that message is projected. The measured causal anatomy directly motivates the first two choices, since attribution is restricted to the read window and each message is scored by its decision-aligned content. The contrast direction answers a separate question, because explaining the realized decision margin and localizing evidence for the predicted class need not favor the same projection. We examine these choices in turn before testing the full design grid.

We first isolate the effect of layer scope while keeping the underlying accounting rule fixed. We reproduce the spatial component of Gandelsman et al. (2024a) under CLIP's native zero-shot readout by taking the component of each position's write into CLS along the predicted-class text direction, after accounting for the realized final-normalization affine and mean-centering the direction as in Eq. (14). On zero-shot-correct images with bounding boxes, restricting this same score from all layers to the read window improves energy pointing by +0.004 [+0.001, +0.008]. The same restriction improves the margin-direction variant by +0.033 [+0.027, +0.038] (Table 8). Thus, the read-window gain is present even when the scoring rule and contrast direction are held fixed.

Table 8: Energy pointing on CLIP under its native zero-shot text head (n=365 zero-shot-correct images with bounding boxes), across layer scope and contrast direction. The single-class all-layer cell reproduces Gandelsman et al. (2024a). Paired comparisons with bootstrap intervals are reported in the text.

Contrast	all layers	read window
single-class $q_{\hat{c}}$	0.868	0.873
margin $q_{\hat{c}} - q_{c_2}$	0.824	0.856
mean-subtracted	0.889	0.910

The contrast direction addresses a different question. The exact map in Eq. (16) attributes the margin between the predicted class and the runner-up, whereas localization asks where evidence for the predicted class is concentrated. At fixed scope, the single-class direction localizes better than the margin direction by +0.016 [+0.004, +0.030]. This is consistent with runner-up evidence often overlapping the predicted object and therefore being removed by the margin contrast.

For localization, we therefore consider a variant that replaces $\tilde{q}$ with the mean-centered single-class direction derived from $q_{\hat{c}}$, while keeping the same read window and realized normalization affine. This improves energy pointing on every backbone in Table 9, with all paired confidence intervals excluding zero. Subtracting the mean class direction produces nearly identical results under the probe readouts, with differences below 0.001. On CLIP's anisotropic text head, it provides an additional localization gain of +0.037 [+0.033, +0.042] over the single-class form while preserving the read-window advantage over all-layer attribution (+0.021 [+0.018, +0.024]).

The margin direction nevertheless remains the form tied to the exact decomposition in Eq. (16). The published single-class all-layer baseline differs from our margin map in both scope and contrast, so comparing the two directly does not isolate the effect of the read window. Indeed, our margin read-window map is statistically indistinguishable from that baseline at -0.012 [-0.026, +0.001], whereas holding the contrast fixed and restricting the margin form to the read window improves localization as shown above. The margin and single-class variants therefore share the same read-path construction but answer different attribution questions. The margin form decomposes the realized decision margin, while the single-class form is better suited to spatial localization.

Table 9: Energy pointing for contrast variants evaluated at the read window under the probe readout (n=394–420 per model). The margin row corresponds to the map in Table 3, while the single-class row is the localization variant. CLIP under its native text head is reported in the text.

Contrast	DINOv2	DINOv3	sup. ViT	DeiT
margin $q_{\hat{c}} - q_{c_2}$	0.887	0.917	0.865	0.864
single-class $q_{\hat{c}}$	0.903	0.927	0.884	0.894

Having isolated the scope effect and the role of the contrast direction, we next test the two anatomy-derived structural choices jointly across all five backbones. These are where attribution is accumulated and what quantity is scored within each attention message. Table 10 crosses layer scope with three weighting rules: attention alone (A), norm-weighted attention ($A\|v\|$) (Kobayashi et al., 2020), and decision-aligned content ($A \times$ content). Every cell uses n=500 images per model, donor shifts $\{1, 7\}$, and a paired bootstrap with 10^4 resamples.

The full grid separates the contributions of the two choices. At fixed weighting, restricting $A \times$ content from all layers to the measured read window improves localization across the five models. At fixed scope, replacing decision-aligned content with raw attention or norm-weighted attention reduces localization. *The anatomy-predicted combination of measured read-window scope and decision-aligned content is the highest-scoring cell in every model*, while single mid-layer maps perform substantially worse and serve as negative controls.

Finally, we ask whether the intervention-defined read-window boundary is also reflected in localization. Extending the $A \times$ content map one layer below the measured window does not improve energy pointing (Table 11). For single-read models, this extends the scope from L11 to L10–11, while for dual-read models it extends L10–11 to L9–11. Thus, the first layer beyond the causally identified window provides no additional localization benefit.

The complementary comparison asks whether a measured read layer should be included when it is present. On supervised ViT and DeiT, whose read window spans L10–11, including both read layers improves localization relative to L11 alone. On the single-read models, adding L10 provides no benefit. The localization boundary therefore mirrors the single-read versus dual-read distinction identified independently by the causal interventions.

Fig. 16 provides qualitative examples of the same design effects. Together, the quantitative and qualitative results support the two structural choices derived from the causal anatomy, decision-aligned message content and a layer scope matched to the measured read window.

H.3 Causal verification and intervention dependence

We first test whether the highest-scoring patches identify content whose replacement causally damages the prediction. Fig. 17 shows graded deletion under fixed-weight resampling at the read window, and Table 12 reports the top-8 result. *Our map exceeds the pre-specified* $5\times$ *random-ranking*

Table 10: Full attention-map design grid, reported as energy pointing across layer scope and weighting rule for five backbones. Weighting rules are attention alone (A), norm-weighted attention ($A\|v\|$) (Kobayashi et al., 2020), and decision-aligned content ($A \times$ content). Shaded cells mark the anatomy-driven designs, while bold and underlined values indicate the best and second-best results, respectively.

Scope	Weight	DINOv2	DINOv3	CLIP	sup. ViT	DeiT
L11	A	0.841	0.848	0.642	0.644	0.578
L11	$A\|v\|$	0.839	0.850	0.743	0.687	0.665
L11	$A \times$ content	**0.887**	**0.917**	**0.859**	0.847	<u>0.847</u>
L10–11	A	0.830	0.800	0.637	0.694	0.597
L10–11	$A\|v\|$	0.829	0.807	0.738	0.748	0.694
L10–11	$A \times$ content	<u>0.885</u>	<u>0.902</u>	<u>0.856</u>	**0.865**	**0.864**
all	A	0.745	0.735	0.644	0.692	0.672
all	$A\|v\|$	0.752	0.742	0.696	0.740	0.725
all	$A \times$ content	0.876	0.893	0.825	<u>0.857</u>	0.845
mid (L6)	A	0.720	0.694	0.631	0.700	0.721
mid (L6)	$A\|v\|$	0.722	0.696	0.655	0.716	0.731
mid (L6)	$A \times$ content	0.738	0.694	0.663	0.728	0.734
random map	—	0.671	0.677	0.671	0.679	0.678

Table 11: Paired energy-pointing differences from widening the $A \times$ content scope, with bootstrap intervals in brackets. The left column extends the scope one layer below the read window, while the right column adds the second read layer. Positive values favor the wider scope.

Model	one layer below window	second read layer added
DINOv2	-0.003 [-0.005, -0.001]	-0.003 [-0.004, -0.001]
DINOv3	-0.015 [-0.018, -0.012]	-0.015 [-0.018, -0.012]
CLIP	-0.003 [-0.007, +0.001]	-0.003 [-0.007, +0.001]
sup. ViT	-0.001 [-0.003, +0.001]	+0.017 [+0.013, +0.022]
DeiT	+0.001 [-0.002, +0.003]	+0.017 [+0.011, +0.023]

criterion in every model, with all bootstrap lower bounds also above the threshold, showing that its ranking is causally informative under the path-matched intervention. This criterion is not unique to our method. GAE (Chefer et al., 2021a) also passes, with ratios 16.6/10.2/13.1/12.8$\times$ across the four evaluated models, and raw attention reaches 9.9$\times$ on DINOv2. The test therefore establishes causal informativeness relative to random rather than uniquely separating our map from other strong attribution methods.

Table 12: Top-1 damage (pp) from replacing the top-8 patches under fixed-weight resampling at the read window (n=500, donor shifts $\{1, 7\}$). The ratio compares our map with a random-8 ranking, with the bootstrap 2.5% lower bound shown in brackets.

Model	ours	all-layer DLA	random-8	ratio [lower]
DINOv2	15.6	15.6	0.9	16.4 [9.6]
DINOv3	31.8	30.3	2.1	15.1 [10.5]
CLIP	26.3	22.6	2.9	9.1 [6.7]
sup. ViT	18.1	17.5	1.2	15.1 [9.2]
DeiT	12.5	10.7	1.6	7.8 [5.2]

Repeating the comparison for $m \in \{1, 2, 4, 8\}$ tests whether the measured layer scope also improves causal selectivity relative to all-layer DLA. The top-m sets selected by the two maps overlap only partially, with mean per-image overlap from 0.53 to 0.88 across models and values of m. Under replacement at the read window, CLIP favors our map at every value of m, with gaps from $+1.3$ to $+3.7$pp and confidence intervals excluding zero, while DINOv3 and DeiT separate at m=8.

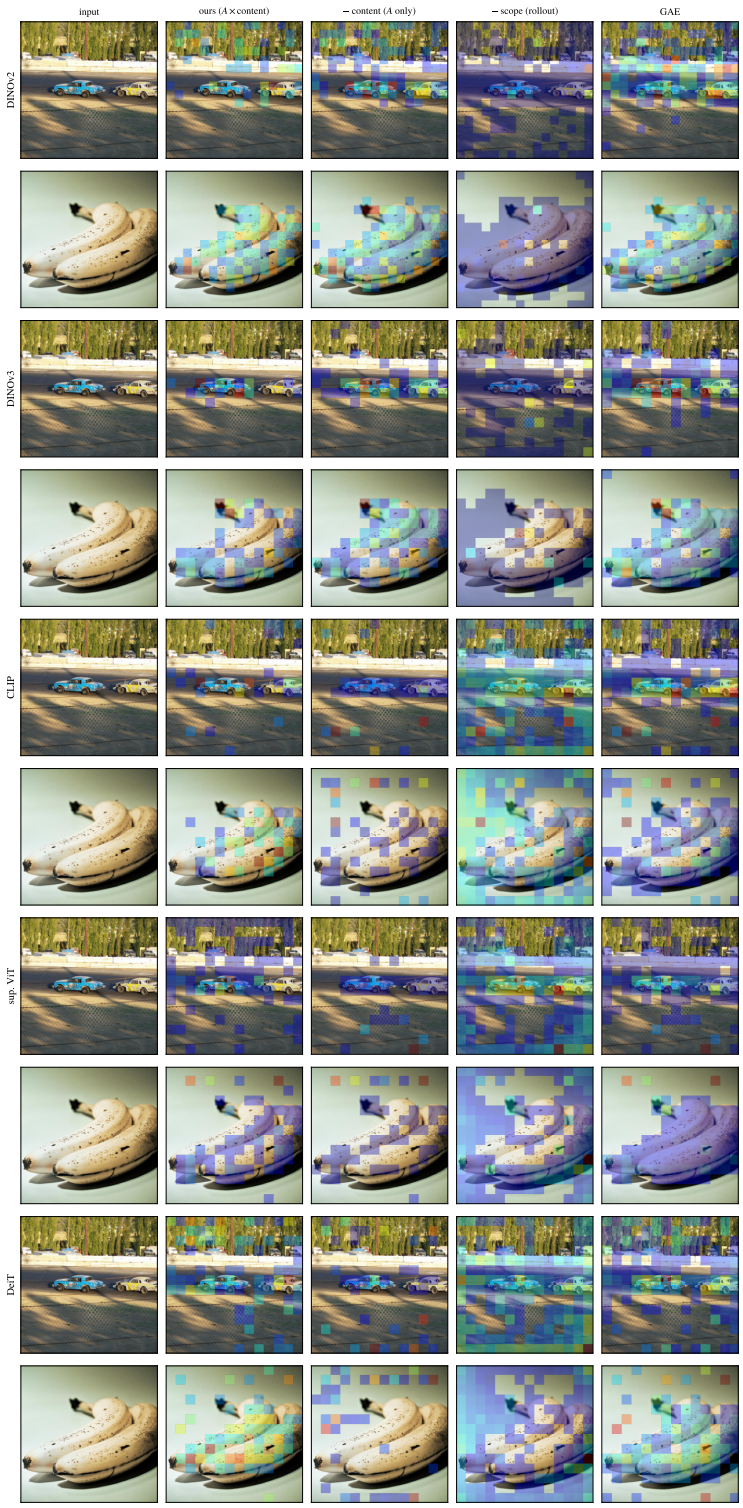


Figure 16: Qualitative view of the design grid. Each panel shows the cells containing the top 80% of positive map mass for the same two inputs across all backbones. Removing the content projection admits more background regions, while broadening the layer scope produces more diffuse maps. GAE remains class-specific but is generally less spatially concentrated than the read-window map.

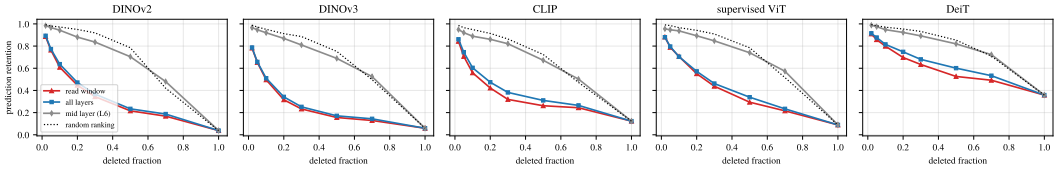


Figure 17: Graded causal deletion under the matched operator. Curves show the fraction of the clean prediction that remains as top-ranked patches are progressively replaced, comparing the $A \times$ content map across layer scopes with a random ranking. Lower prediction retention indicates greater causal damage.

The same comparison at the image input shows no consistent advantage, with seventeen of the twenty model-by-m comparisons not significant and the isolated differences pointing in opposite directions on DINOv2 and supervised ViT. This contrast reflects the intervention semantics. Read-window replacement directly perturbs the patch content attributed by our map, whereas input-level replacement affects every downstream route originating from that patch. The path-matched intervention therefore provides the more direct test of the specific attribution claim made by the read-window map, while the input-space intervention evaluates a broader notion of patch importance.

This operator dependence also appears in the broader method comparison. Norm-weighted attention $A\|v\|$ is competitive with GAE on causal deletion AUC in four of five models. On DINOv2, for example, it reaches 0.22, compared with 0.28 for our map and 0.22 for GAE. This is consistent with the matched operator swapping value content while holding attention fixed, which makes value norm a strong predictor of damage even without class direction. Under input-space deletion, this advantage largely disappears and the ordering reverses in four of five models. *Deletion rankings therefore depend not only on the attribution map but also on what the intervention removes.*

Across method families, the comparison is mixed. The read-window map improves over its mid-layer and all-layer counterparts, but GAE is generally stronger on deletion faithfulness, while our map achieves higher energy localization in Table 3. This is consistent with the decomposition in Fig. 15, since our map attributes only the patch-delivered component of the read whereas gradient-based methods can also propagate through other paths such as read-layer MLPs. The proposed map should therefore not be interpreted as uniformly superior across saliency criteria. Its main distinction is that its scope and scoring rule are derived from the measured causal anatomy.

This appendix separates three claims about the attribution map. The exact decomposition specifies which component of the realized margin is assigned to patches. The design analysis shows that both the measured layer scope and decision-aligned content improve localization. The intervention experiments then show that high-scoring patches are causally informative under a path-matched replacement. The map should therefore be interpreted as an anatomy-derived attribution of the terminal attention read, rather than as a complete decomposition of the classification decision or a uniformly superior saliency method across evaluation criteria.

I Full saliency benchmark

This appendix provides the full saliency benchmark summarized in §5.3, testing the read-window map across complementary notions of faithfulness and localization, including evaluations that do not use the intervention from which the map was derived. We first evaluate deletion faithfulness under matched and operator-neutral interventions (§I.1). We then evaluate localization and separate genuine class-specific signal from spatial concentration and coverage effects (§I.2). Finally, we test external validity on ImageNet-S segmentation masks (§I.3).

We compare our DLA map with raw last-layer CLS attention, attention rollout (Abnar & Zuidema, 2020), GAE (Chefer et al., 2021a), LeGrad-style and GMAR adaptations (Bousselham et al., 2025; Jo et al., 2025), the per-patch logit lens (Schaffer et al., 2025), and a causal head-profile ablation. The head-profile baseline weights each attention head by its causal damage measured on disjoint images, with the resulting profiles shown in Fig. 18c. Because it uses causal information unavailable to the other methods, we treat it as an ablation rather than an independent baseline.

We evaluate deletion under two operators and localization under two metrics. The first deletion benchmark uses the fixed-weight replacement operator matched to the semantics of our map, while the second replaces patches at the image input and provides an operator-neutral comparison. Localization is evaluated on correctly predicted images against ILSVRC bounding boxes using argmax pointing and energy pointing. Argmax comparisons use exact McNemar tests on per-image outcomes, while energy comparisons use paired bootstrap confidence intervals with 10,000 resamples.

We do not include full relevance-propagation methods such as Chefer-LRP (Chefer et al., 2021b), TokenTM (Wu et al., 2024), and AttnLRP (Achtibat et al., 2024), which require method-specific propagation stacks across the heterogeneous model implementations considered here. Our LeGrad-style and GMAR adaptations instead admit closed-form implementations in $(A, \nabla A)$. This omission limits the completeness of the benchmark. All per-model results are reported in Table 13.

I.1 Deletion faithfulness

Table 13: Full saliency benchmark across five backbones. Arrows give the preferred direction of each metric. Deletion AUC is reported under fixed-weight causal replacement and input-space patch replacement. Pointing and energy localization are evaluated on correctly predicted images (409/394/362/401/421 images per model). Causal deletion for the supervised models uses the L11-only operator, with the scope-matched two-layer comparison reported in the text.

Model	Metric	Ours	Raw attn	Rollout	GAE	LeGrad	GMAR	Patch-lens	Head-prof	Random
DINOv2	AUC-causal (↓)	0.296	0.239	0.474	0.231	0.403	0.475	0.391	0.238	0.650
	AUC-input (↓)	0.240	0.223	0.304	0.196	0.286	0.303	0.287	0.222	0.411
	pointing (↑)	0.958	0.922	0.733	0.966	0.873	0.743	0.888	0.924	0.660
	energy (↑)	0.887	0.841	0.696	0.814	0.685	0.698	0.770	0.840	0.669
DINOv3	AUC-causal (↓)	0.232	0.263	0.485	0.240	0.351	0.505	0.398	0.261	0.656
	AUC-input (↓)	0.184	0.213	0.295	0.175	0.239	0.301	0.270	0.208	0.396
	pointing (↑)	0.954	0.919	0.769	0.959	0.916	0.741	0.881	0.914	0.704
	energy (↑)	0.917	0.848	0.716	0.822	0.701	0.695	0.773	0.854	0.679
CLIP	AUC-causal (↓)	0.323	0.304	0.497	0.285	0.399	0.494	0.706	0.302	0.647
	AUC-input (↓)	0.179	0.204	0.270	0.165	0.181	0.268	0.353	0.204	0.312
	pointing (↑)	0.917	0.533	0.555	0.762	0.925	0.558	0.688	0.536	0.682
	energy (↑)	0.859	0.642	0.667	0.756	0.706	0.668	0.654	0.659	0.670
sup. ViT	AUC-causal (↓)	0.705	0.683	0.780	0.690	0.800	0.778	0.776	0.683	0.840
	AUC-input (↓)	0.224	0.306	0.344	0.209	0.233	0.334	0.313	0.305	0.435
	pointing (↑)	0.873	0.584	0.743	0.673	0.945	0.776	0.843	0.584	0.698
	energy (↑)	0.865	0.644	0.690	0.774	0.703	0.692	0.710	0.645	0.678
DeiT	AUC-causal (↓)	0.847	0.845	0.897	0.850	0.898	0.890	0.882	0.846	0.918
	AUC-input (↓)	0.252	0.360	0.415	0.236	0.257	0.416	0.311	0.360	0.440
	pointing (↑)	0.922	0.470	0.634	0.743	0.850	0.590	0.848	0.468	0.688
	energy (↑)	0.864	0.578	0.673	0.760	0.702	0.672	0.732	0.577	0.677

GAE performs best overall on deletion faithfulness in Table 13. Under the matched fixed-weight operator, it leads on DINOv2 and CLIP and is statistically tied with our map on DINOv3 (-0.008 [-0.022, +0.006]). *The read-window map therefore does not uniformly dominate existing attribution methods under causal deletion.*

The supervised models require a scope-matched deletion operator because their read window spans two layers. With an L11-only intervention, even a random ranking retains 0.57 on supervised ViT and 0.79 on DeiT at full deletion, leaving little room to separate methods. We therefore also evaluate a matched two-layer intervention. Under this operator, the read-window map reaches 0.364 on supervised ViT and 0.568 on DeiT, compared with 0.325 and 0.542 for GAE. GAE remains stronger, but matching the intervention to the measured read window substantially increases the discriminative power of the benchmark.

The same broad ordering persists under operator-neutral input-space deletion. GAE remains the strongest baseline overall, while the read-window map remains competitive. The main exception is supervised ViT, where LeGrad-style slightly outperforms the terminal-layer variant of our map

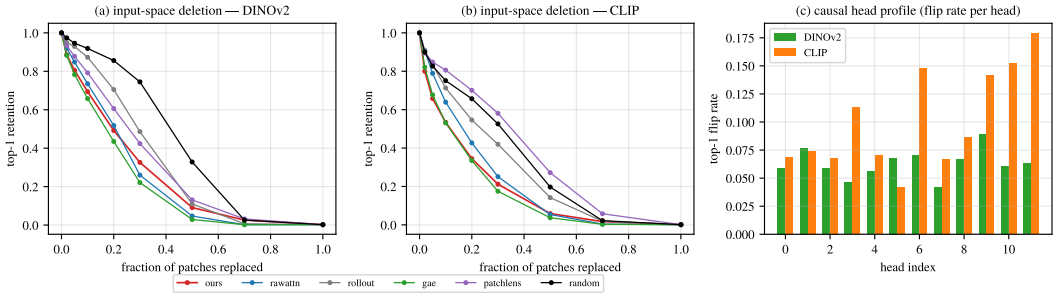


Figure 18: Saliency benchmark diagnostics. **(a,b)** Operator-neutral input-space deletion curves on DINOv2 and CLIP. **(c)** Per-head causal profiles for the terminal read, which is head-differentiated on CLIP and broader on DINOv2.

(0.243), though not the read-window variant. Representative input-space deletion curves are shown in Fig. 18a,b.

The deletion results are also stable across several perturbation choices. Donor, blur, and per-image-mean imputations preserve the same broad ordering, addressing imputation-sensitivity concerns (Rong et al., 2022). On DINOv2, our single-donor scores are 0.222/0.351/0.304, compared with the two-donor benchmark value 0.240. Insertion gives a similar comparison (Petsiuk et al., 2018), with our map reaching 0.368 and GAE reaching 0.392, where higher values are better. The LeGrad-style and GMAR adaptations trail our map and GAE on causal deletion across all five models.

I.2 Localization, concentration, and class specificity

We evaluate localization with argmax pointing (Zhang et al., 2018) and energy pointing (Wang et al., 2020). Argmax pointing tests whether the highest-scoring patch falls inside the object box, whereas energy pointing measures the fraction of positive attribution mass contained within it. Argmax pointing is strongly ceiling-compressed in our setting, since object boxes cover 0.67 of the crop in expectation and even a fixed center patch reaches 0.89 on DINOv2 and 0.89 on CLIP. It is also computed on model-specific image subsets and is not directly comparable across models. We therefore emphasize energy pointing as the more discriminative localization measure.

Our map achieves higher energy pointing than GAE across all five models, with paired confidence intervals excluding zero. Argmax pointing gives a less consistent comparison. Our map outperforms GAE on CLIP ($p = 5.8 \times 10^{-12}$), while the DINOv2 and DINOv3 comparisons are ceiling-level ties ($p = 0.607$ and $p = 0.824$). On supervised ViT, LeGrad-style achieves a higher argmax score despite a more diffuse energy map. The two metrics therefore capture different properties of the attribution map rather than interchangeable notions of localization.

Bounding-box coverage can itself inflate energy scores. We therefore also report the normalized quantity $(E - \mathrm{cov})/(1 - \mathrm{cov})$, rescaling the energy score E by the box coverage cov, on images with $\mathrm{cov} < 1$, which comprise about 80% of each pointing subset. Our normalized scores are 0.70/0.78/0.58/0.55 compared with 0.47/0.46/0.25/0.31 for GAE. The energy advantage is therefore not explained solely by bounding-box coverage.

Energy pointing also rewards spatial concentration, and our map is approximately 2–3 times more concentrated than GAE. The cells carrying 80% of positive mass span 33–50 for our map and 93–124 for GAE, as illustrated in Fig. 19. To separate concentration from localization quality, we match the sparsity of GAE to that of our map by retaining the same number of cells on each image and renormalizing the remaining mass. Under this control, the energy gap reverses on the self-supervised models, with -0.037 [-0.045, -0.029] on DINOv2 and -0.021 on DINOv3, becomes marginal on supervised ViT at +0.010 [+0.000, +0.019], and remains positive on CLIP and DeiT at +0.028/+0.031. *Much of the raw energy advantage on the self-supervised models therefore comes from stronger spatial concentration*, while a localization advantage remains on CLIP and the supervised models after matching sparsity. The within-method design-grid comparisons of §5.2 are unaffected by this control.

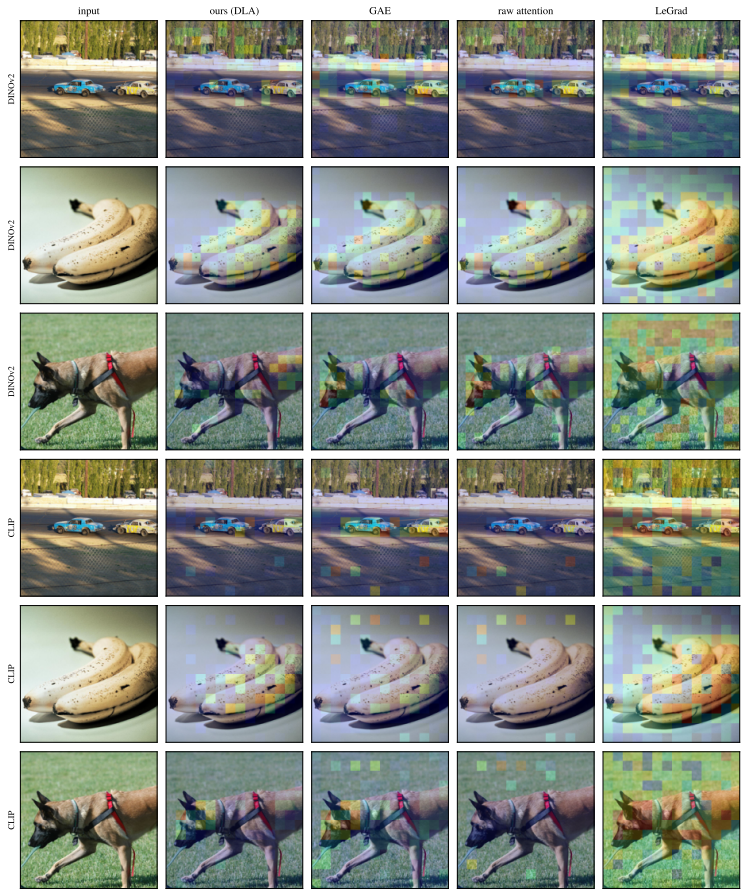


Figure 19: Representative predicted-class attribution maps on DINOv2 and CLIP. The read-window map is more spatially concentrated, while GAE and the other baselines generally distribute attribution over broader regions.

We next test class specificity through parameter randomization (Adebayo et al., 2018). Randomizing the probe reduces pointing from 0.96 to 0.90 and moves deletion from 0.240 toward the random value 0.317, supporting the presence of a class-specific component. At the same time, substantial class-agnostic spatial structure remains under the randomized head, with energy scores of 0.83/0.84/0.73/0.72. On DINOv2 and DINOv3, this residual structure alone still localizes more strongly than GAE, whereas on CLIP and supervised ViT a larger part of the localization advantage depends on the class-specific signal. The localization results should therefore be interpreted as combining generic spatial salience with class-specific margin modulation, rather than as purely class-specific attribution.

I.3 External validity on ImageNet-S

Bounding boxes include background pixels and primarily reward spatial concentration, so we repeat the localization evaluation using segmentation masks from 300 ImageNet-S919 validation images (Gao et al., 2022), one image per category, with each map targeted to the model's own predicted class. We report three complementary metrics. Mask-energy measures the fraction of positive attribution mass inside the foreground mask. Pixel AP is threshold-free, while foreground IoU uses Otsu thresholding, and both place greater emphasis on recovering the full spatial extent of the object. A uniform map has expected mask-energy equal to the foreground coverage 0.41.

Our map achieves higher mask-energy than GAE on all three evaluated models (Table 14), showing that its concentration advantage persists with segmentation masks rather than bounding boxes. Metrics that reward broader object coverage give a more mixed result. On pixel AP, GAE leads on DINOv2,

Table 14: ImageNet-S919 segmentation localization (n=300 validation images, one per category), with each map targeted to the model's own predicted class. Mask-energy is the fraction of positive attribution mass inside the object mask; pixel AP is threshold-free; foreground IoU uses Otsu thresholding. The Δ column gives the paired ours-minus-GAE mask-energy difference with its bootstrap interval.

	mask-energy			pixel AP		foreground IoU	
Model	ours	GAE	Δ [CI]	ours	GAE	ours	GAE
DINOv2	0.70	0.59	0.11 [0.09, 0.12]	0.68	0.78	0.30	0.38
CLIP	0.62	0.51	0.11 [0.09, 0.13]	0.63	0.64	0.29	0.27
sup. ViT	0.64	0.53	0.10 [0.09, 0.12]	0.70	0.69	0.37	0.26

the methods are close on CLIP, and our map leads on supervised ViT. On foreground IoU, our map leads on CLIP and supervised ViT but trails on DINOv2. The DINOv2 result makes the tradeoff especially clear, since our map leads on mask-energy while trailing on both extent-oriented metrics. *The concentration versus coverage tradeoff therefore persists under a tighter spatial ground truth and a separate dataset split.*

Taken together, the benchmark does not show a uniform advantage of the read-window map across saliency criteria. GAE is generally stronger on deletion faithfulness, while the read-window map is more spatially concentrated and achieves higher energy localization. Part of this localization advantage, especially on the self-supervised models, is explained by sparsity, and extent-oriented segmentation metrics give a more mixed comparison. The contribution is therefore not a universally stronger saliency method, but an attribution map whose layer scope and scoring rule are derived from the measured causal path and remain competitive across complementary evaluations of faithfulness and localization.