SOLA: A Structured Operator Library for Attention in Pretrained Vision Transformers

Enzo Tartaglione
LTCI, Télécom Paris, Institut Polytechnique de Paris, France
ISASI-CNR, Italy
enzo.tartaglione@telecom-paris.fr

Vito Paolo Pastore
MaLGa-DIBRIS, University of Genoa, Italy
AIGO, Istituto Italiano di Tecnologia, Italy
vito.paolo.pastore@unige.it

Abstract

Vision Transformers dominate modern image benchmarks; however, what trained self-attention computes head by head remains unclear: prior work either injects inductive biases at initialization or clusters attention patterns visually without committing to a closed-form computation. We ask whether the softmax$(QK^\top)$ inside a trained head can be replaced, without further training, by a structured closed-form operator while preserving prediction and the residual-stream trajectory. We answer yes across ten ViTs and seven pretrainings (ImageNet-1k, ImageNet-21k FT, MAE, DINO, DINOv2, CLIP, SigLIP), and package the result as SOLA, a Structured Operator Library for Attention with three entries, each gated by a per-head diagnostic that upper-bounds substitution error. The entries are a fixed 2D convolution kernel for shift heads, a per-image broadcast row for fixed-target heads, and a per-image rank-k SVD of the attention matrix for near-global heads. A rank-adaptive extension picks k per head from the SVD spectrum and extends structural coverage to every head on every backbone, with task-fidelity controlled by a single spectral threshold τ that trades compression for downstream fidelity. Structural coverage is distinct from verified losslessness: at $\tau{=}0.9$ the full substitution costs 0.9-2.2% top-1 on five supervised classifiers and keeps CLS cosine similarity at or above 0.96 on MAE, DINO, and CLIP, while DINOv2 and SigLIP require a stricter τ; the heads that are both compressed and verified lossless by a standalone test range from 3% (SigLIP) to 84% (MAE) of a backbone. The diagnostics carry over to cross-attention: every decoder cross-attention head in BLIP captioning has effective rank ≈ 1 and admits the broadcast substitution. The code base is available at https://github.com/enzotarta/SOLA.

1 Introduction

Transformers have become the dominant architecture across language, vision, and multimodal learning, and Vision Transformers (ViTs) [9, 25] match or surpass convolutional networks on standard image benchmarks. Their core operation is multi-head self-attention: a softmax over $QK^\top$ followed by a value aggregation. It is in principle expressive enough to implement any content- and position-dependent linear aggregation of input tokens. Cordonnier et al. [5] formalize one face of this expressivity: a multi-head self-attention layer with relative position encodings can implement any $K \times K$ 2D convolution exactly. Self-attention is therefore strictly more expressive than convolution; the question of what trained ViTs actually compute, however, remains open.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Mechanistic interpretability seeks to read computation *out* of trained models rather than reason from architecture alone. In language, this program has identified narrow but specific algorithms: induction heads [19], the indirect-object circuit in GPT-2 [30], and a broader catalogue of attention-mediated copy/move/compare operations underwritten by a mathematical framework for transformer circuits [11]. Recent work on monosemanticity and dictionary learning of MLP features [1] extends this to non-attention components. In vision, the analogous picture is sparser. Self-attention maps have been visualized for individual heads [2, 6] and clustered into recurring patterns by visual analytics [33, 34], but the descriptions are limited to qualitative groupings.

Existing approaches to ViT interpretability fall into two camps. The *constructive* camp injects convolutional or local inductive biases at initialization or architecture, ConViT [10], Transformed CNNs [7], Mimetic Initialization [26], Swin's shifted windows [17], deformable attention [32], and Shift-T's [29] replacement of attention by token shift; these papers prove what is *possible* but say nothing about what trained models *do*. The *read-out* camp groups trained heads into recurring attention patterns by visual clustering [33, 34] or compares feature representations across pretraining objectives [2, 28]; these stop short of committing to a closed-form operator per cluster, and do not test whether replacing a head with that operator preserves the model's predictions. Both camps miss the operational claim: *can the softmax$(QK^\top)$ in a trained head be replaced*, training-free and head-by-head, with a structured closed-form computation, while preserving prediction, per-class behavior, and the residual-stream trajectory?

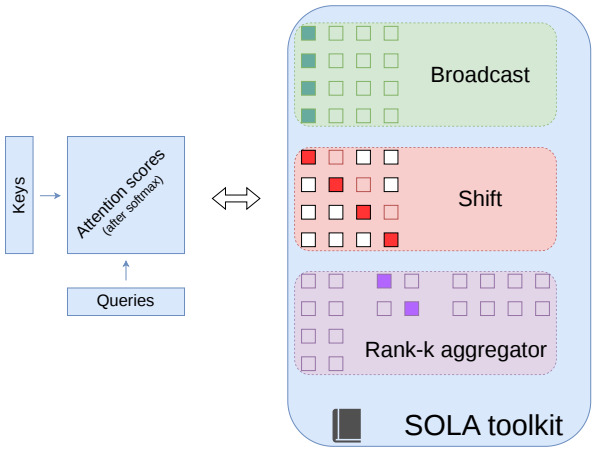

Figure 1: The softmax$(QK^\top)$ attention-score matrix of a trained head is replaced, in place, by one of three closed-form operators: BROADCAST, SHIFT, and RANK-k AGGREGATOR.

We give an affirmative answer across ten pretrained ViTs and seven pretrainings, and package the result as **SOLA** (Figure 1), a Structured Operator Library for Attention with three closed-form substitutes for the softmax$(QK^\top)$ score matrix, gated by per-head diagnostics and applied at the level of the attention scores so that the value projection V is left untouched. (i) *Shift heads* are 2D translations, replaceable by a fixed 2D-convolution aggregator on V that bypasses the Q, K, softmax pathway entirely. (ii) *Broadcast heads* compute one image-dependent global summary of V and distribute it identically to all queries; we substitute them with a per-image, query-independent broadcast of the attention row-mean. (iii) *Rank-k aggregators* are query-dependent but low effective rank, well approximated by a per-image rank-k SVD of the attention matrix $A^{(h)}$ with k matched to the head's effective rank. A rank-adaptive extension (Cor. 3.1) extends structural coverage to every head on every backbone, with task-fidelity controlled by a single spectral threshold τ. The per-layer mixture is set by the pretraining objective: MAE sharpens shifts; SigLIP loads the fixed-target category; DINO and DINOv2 spread mass through near-global. Each per-operator substitution, applied at its diagnostic-gated heads, is functionally interchangeable with the head it replaces, preserving predictions within the 95% CI of zero on full ImageNet val, residual-stream fidelity at every block boundary, and downstream attention outputs under path patching. Our contributions are the following:

- **SOLA**: three training-free, per-head, drop-in substitutions for softmax$(QK^\top)$ on frozen pretrained ViTs. A fixed 2D convolution kernel for shift heads, a per-image query-independent broadcast for fixed-target heads, and a rank-k SVD aggregator for near-global and diffuse-local heads; a rank-adaptive extension (Cor. 3.1) sets k per head from the SVD spectrum and gives full structural coverage on every backbone, with its fidelity cost reported per backbone (Section 3.2; per-backbone breakdown in Table 2).
- An operator taxonomy and per-head diagnostic battery covering eight structural categories, with each diagnostic (recentered L_1 residual, i.e., how far each attention row is from one fixed spatial ker-

nel; full-attention L_1 residual, i.e., how far the rows are from their per-image mean; effective rank, i.e., how many singular directions carry most of the energy of the attention matrix) quantitatively upper-bounding the error of the substitution it admits (Section 3.3; Propositions 1-3).

- Cross-pretraining structure across ten pretrained ViTs and seven pretrainings (ImageNet-1k, ImageNet-21k FT, MAE, DINO, DINOv2, CLIP, SigLIP): shift counts and sharpness, fixed-target population size, and near-global effective rank vary systematically with the pretraining objective (Section 4.1), extending to substitutable operators the spectral differences between objectives reported by Doshi et al. [8].

Low-rank attention and convolution-like heads are known (Sec. 2); our contribution is the closed-form substitution of each operator with a computable per-head error bound, the taxonomy assigning every head to one of them, and the test of the bound against prediction preservation. To our knowledge, this is the first work to map every softmax$(QK^\top)$ head in frozen pretrained ViTs, head-by-head, to a *closed-form, training-free* structured operator from a small fixed inventory and test these substitutions for prediction preservation across multiple backbones. We keep apart *structural coverage* (an entry with a computable error bound exists; 100% of heads), *compression* ($k_h \leq d_h$; 93.4% at τ=0.9), and *verified losslessness* (a standalone test preserves predictions; 2.8% to 84.0% of the heads of a backbone, Table 2). Prior post-hoc replacements either require distillation or fine-tuning [23, 24], use a single global learned operator class [3, 13, 31], or stop at structural analysis without a substitution test [8, 34].

2 Related work

Attention's expressive power and convolutional inductive biases. On the theory side, Cordonnier et al. [5] show an attention layer with $H \geq K^2$ heads and relative position encodings can implement any $K \times K$ convolution by collapsing each head to a one-hot at one offset; the ViTs we study use absolute position embeddings, so the construction holds only approximately. On the architectural side, locality is injected at training time: ConViT [10], Transformed CNNs [7], and Mimetic Initialization [26] bias the initialisation; Swin [17] restricts attention to shifted local windows; Deformable Attention [32] predicts a sparse content-dependent key set per query. These works prove what is *possible* or what an architectural prior buys at training time, but say nothing about what trained vanilla self-attention *is*. We read the operator out of frozen pretrained attention, head-by-head, with no architectural modification.

Empirical head characterisation: pruning, taxonomies, and replacements. A complementary line measures head behaviour in trained transformers. Pruning studies show that most attention heads can be removed with little loss in language models [18, 27]. In vision, AttentionViz [34] and Yamamoto et al. [33] cluster trained ViT attention to recurring visual patterns and compare them to human eye-tracking; Caron et al. [2] highlight emergent object-segmentation attention in self-supervised ViTs; Darcet et al. [6] attribute spurious high-norm late-layer tokens to register tokens. The closest prior work is Doshi et al. [8], who apply an SVD to the query-key weight product $W_Q W_K^\top$ of each head; the observations that heads specialize, that the SVD spectrum reveals this structure, and that DINOv2 has a flatter spectrum than supervised ViTs are theirs. Their analysis is descriptive; SOLA substitutes the realized attention matrix and bounds and measures the error. Architecturally, Shift-T [29] replaces *all* attention with a fixed token-shift operation in a network trained from scratch and reports degraded performance. In contrast, our characterisation is quantitative and *operator-resolved*: each trained head is mapped to a closed-form substitute (shift, broadcast, or rank-k) gated by a per-head diagnostic that upper-bounds the substitution error. Trained attention already contains literal shifts in a small subset of heads, broadcasts in another, and rank-bounded aggregators in a third, a hybrid mix rather than the single universal operator that Shift-T's negative result rules out.

Mechanistic interpretability. The transformer-circuits programme formalises attention as a sum of linear operators on the residual stream [11] and identifies specific components: induction heads [19], the indirect-object circuit in GPT-2 [30], automatic circuit discovery [4], and dictionary-learning decompositions of MLP activations [1]. These works characterise attention through the *role* of a head in a larger circuit, established via causal interventions and patch-level analysis. Query-key interactions have also been decomposed in language models [15], and the positional bias of DINO-family ViTs reduced by fine-tuning with ALiBi encodings [21]; neither replaces heads post hoc. Our claim is complementary and operates one abstraction level lower: at the per-head level, the softmax$(QK^\top)$ output of trained vision attention is captured by a closed-form classical operator

drawn from a small fixed library, with no training and no circuit-level reasoning required for the per-head step. A path-patching probe checks that our substitutions preserve the downstream circuit they participate in.

Low-rank, structured, and post-hoc replacement attention. A long line exploits low-rank or structured priors at training time: Linformer [31] projects keys and values into a fixed low-dimensional subspace; Performer [3] approximates softmax attention with random feature kernels; Han et al. [13] model attention as block-circulant-with-circulant-blocks matrices to obtain $O(N \log N)$ complexity. A more recent line proposes *post-hoc* replacements of attention in pretrained models, but each requires training to recover function: FAR [23] replaces all self-attention in pretrained DeiTs with a multi-head bidirectional LSTM via block-wise distillation; LUNA [24] swaps softmax for a learned linear kernel and fine-tunes briefly on BERT/ViT. We make the substitution training-free and operator-resolved: the rank-k entry of SOLA is a per-image, post-hoc *measurement* on the trained model: the rank threshold needed for the substitution-error bound is read off the SVD spectrum rather than imposed at training time, and the rank-adaptive extension picks k per head from the spectrum rather than fixing a global subspace dimension or learning a replacement module.

Pretraining objectives, downstream test beds, and shape-vs-texture. We sweep ten ViTs spanning supervised (DeiT [25], ViT-B ImageNet-21k→ImageNet-1k [9]), masked image modelling (MAE [14]), self-distillation (DINO [2], DINOv2 [20]), and image-text contrastive (CLIP [22], SigLIP [35]) families. Walmer et al. [28] compare supervised, MoCo, MAE, DINO, and CLIP ViTs at the level of representations and saliency; we complement that representation-level comparison with an operator-level inventory and show that the training signal, more than the architecture or the data alone, sets the inventory (two ViT-Bs trained under the same supervised objective on different data give nearly identical taxonomies). Liu et al. [16] build LLaVA on a CLIP vision tower, giving a downstream multimodal test bed. Geirhos et al. [12] show ImageNet-trained CNNs are biased toward texture, and ViTs partially correct this; we go one level deeper, describing the operator inventory through which attention computes and how it varies with pretraining signal.

3 Method

3.1 Setup and notation

Let f be a pretrained ViT on a spatial grid of $N = H \times W$ patch tokens (plus a CLS token and possibly register tokens, which we strip before locality analysis). At each layer ℓ, head h computes

$$A^{(\ell,h)}(x) \;=\; \mathrm{softmax}\!\left(\frac{Q^{(\ell,h)}(x)\,K^{(\ell,h)}(x)^{\top}}{\sqrt{d_h}}\right), \qquad o^{(\ell,h)}(x) \;=\; A^{(\ell,h)}(x)\,V^{(\ell,h)}(x), \quad (1)$$

with the softmax taken row-wise. We use the following symbols throughout:

- d, n_h, $d_h = d/n_h$: model embedding dim, heads per layer, per-head dim;
- $\alpha_{i,j}^{(\ell,h)}$: entry (i,j) of $A^{(\ell,h)}$; when discussing a single head, we drop ℓ and write $A^{(h)}$, $\alpha_{i,j}$;
- $\mathcal{A} = \{A \in \mathbb{R}^{N\times N} : A\mathbf{1} = \mathbf{1},\ A_{ij} \geq 0\}$, the row-stochastic matrices, where $\mathbf{1}$ is the all-ones column;
- $\mathbb{1}[\cdot]$ is the indicator; $\|\cdot\|_F,\ \|\cdot\|_{\mathrm{op}},\ \|\cdot\|_p$ are the Frobenius, spectral operator, and standard p-norms.

Operator classes. We test the hypothesis that each trained head's $A^{(h)}$ concentrates near one of three structured subsets of $\mathcal{A}$:

$$\begin{aligned}
\mathcal{A}_{\mathrm{sh}}(\Delta) &= \big\{A \in \mathcal{A} : A_{ij} = \mathbb{1}[j = i + \Delta]\big\} && \text{(shift at offset } \Delta \in \mathbb{Z}^2\text{)},\\
\mathcal{A}_{\mathrm{br}} &= \big\{\mathbf{1}_N\, a^{\top} : a \in \mathbb{R}^N_{\geq 0},\ a^{\top}\mathbf{1} = 1\big\} && \text{(broadcast: every row equals same } a\text{)},\\
\mathcal{A}_{\leq k} &= \big\{A \in \mathcal{A} : \mathrm{rank}(A) \leq k\big\} && \text{(rank-}k\text{ aggregator)}.
\end{aligned}$$

$\mathcal{A}_{\mathrm{br}} \subset \mathcal{A}_{\leq 1}$; $\mathcal{A}_{\mathrm{sh}}(\Delta)$ is a permutation matrix, hence full-rank and disjoint from $\mathcal{A}_{\leq k}$ for $k < N$.

Kernel and offset notation. We use $\delta \in \mathbb{Z}^2$ for a spatial offset on the recentered grid and reserve the symbol $\bar{\kappa} : \mathbb{Z}^2 \to \mathbb{R}_{\geq 0}$ for the head's empirical mean kernel (defined in Proposition 1 below). The shorthand $\alpha_{i,i+\delta}$ denotes the entry of $A^{(h)}$ at query i and key $i + \delta$, treating positions on the

spatial grid as 2D coordinates. We write $o = A^{(h)}V$ for the head output (Eq. 1) and $\hat{o}$ (or o_{br}, $o_{\le k}$) for the corresponding substituted output.

3.2 Substitutions and exactness bounds

We propose three closed-form substitutions, one per operator class of Sec. 3, each the orthogonal projection of $A^{(h)}$ onto the class in question. *All three act on the attention matrix $A^{(h)}$ only; the value projection $V^{(h)} = W_V^{(h)}x$ is left unchanged.* The substitution error is bounded uniformly by the projection distance:

$$\|AV - A'V\|_F \;\le\; \|A - A'\|_F \, \|V\|_{\mathrm{op}}, \tag{2}$$

so each proposition reduces to bounding $\|A - A'\|_F$ via a per-head diagnostic.

Proposition 1 (shift / kernel substitution). *Substitute $A^{(h)}$ with the row-stochastic matrix $\Phi_{\bar{\kappa}}$ induced by the empirical mean kernel $\bar{\kappa}(\delta) = \mathbb{E}_{x,i}[\alpha_{i,i+\delta}]$:*

$$\Phi_{\bar{\kappa}}[i,j] = \bar{\kappa}(j-i), \qquad \text{on-grid renormalised at boundary queries.}$$

The translation-equivariance (TE) residual is the L_1 deviation of each attention row $\alpha_{i,\cdot}$ from the kernel row $\Phi_{\bar{\kappa}}[i,\cdot]$,

$$r = \mathbb{E}_{x,\,i}\left\|\alpha_{i,\cdot} - \Phi_{\bar{\kappa}}[i,\cdot]\right\|_1,$$

averaged over images x and interior queries i. The substituted output $\hat{o} = \Phi_{\bar{\kappa}}V$ satisfies

$$\|o - \hat{o}\|_F \;\le\; \sqrt{2Nr}\,\|V\|_{\mathrm{op}}.$$

Tightness: equality at r=0 when $A^{(h)} \in \mathcal{A}_{\mathrm{sh}}(\hat{\Delta})$ for the fingerprint argmax $\hat{\Delta} = \arg\max_\delta \bar{\kappa}(\delta)$. *Special case*: for $\bar{\kappa}(\delta) = \mathbb{1}[\delta = \Delta]$ (one-hot), $\Phi_{\bar{\kappa}}V_i = V_{i+\Delta}$ exactly, and the $Q, K, QK^\top$ matmul and softmax can all be skipped.

Proposition 2 (per-image broadcast). *For each image x, let $\bar{a}_x \in \mathbb{R}^N$ be the per-image column-mean, $\bar{a}_x[j] = \frac{1}{N}\sum_i \alpha_{i,j}^{(h)}(x)$, and substitute*

$$A_{\mathrm{br}}(x) = \mathbf{1}_N\,\bar{a}_x^\top \quad \text{(every row equal to } \bar{a}_x\text{)}.$$

The image-mean residual

$$r_{\mathrm{abs}} = \mathbb{E}_x\left[\frac{1}{2N}\sum_i \|\alpha_{i,\cdot}^{(h)}(x) - \bar{a}_x\|_1\right]$$

quantifies how much $A^{(h)}(x)$ varies across rows. Then

$$\|o - o_{\mathrm{br}}\|_F \;\le\; 2\sqrt{N\,r_{\mathrm{abs}}}\,\|V\|_{\mathrm{op}}.$$

Tightness: equality at r_{abs}=0, which happens when $A^{(h)}(x) \in \mathcal{A}_{\mathrm{br}}$ for every x.

Proposition 3 (per-image rank-k). *Substitute $A^{(h)}(x)$ with its best rank-k SVD approximation $A^{(h)}_{\le k}(x)$. Let $\sigma_1 \ge \cdots \ge \sigma_N$ be the singular values of $A^{(h)}(x)$, and define the per-image effective rank*

$$k_\tau(x) = \min\left\{k : \textstyle\sum_{i\le k}\sigma_i^2 \ge \tau\,\|A^{(h)}(x)\|_F^2\right\}.$$

For any $k \ge k_\tau(x)$,

$$\|o - o_{\le k}\|_F \;\le\; \sqrt{1-\tau}\,\|A^{(h)}(x)\|_F\,\|V\|_{\mathrm{op}}.$$

Corollary 3.1 (adaptive per-head rank-k). *Let $S \subseteq [n_h]$ be the substituted heads and set*

$$k_h = \left\lceil\, \mathbb{E}_{x\in\mathcal{C}}\, k_\tau^{(h)}(x)\,\right\rceil \quad \text{on a calibration batch } \mathcal{C}.$$

Substituting every $h \in S$ by $A^{(h)}_{\le k_h}$, the per-head bound of Proposition 3 holds with $k = k_h$ whenever $k_\tau^{(h)}(x) \le k_h$, and the layer-level residual-stream error aggregates linearly:

$$\left\|\textstyle\sum_{h\in S}\left(o^{(h)} - o^{(h)}_{\le k_h}\right)\right\|_F \;\le\; \sqrt{1-\tau}\,\textstyle\sum_{h\in S}\|A^{(h)}(x)\|_F\,\|V^{(h)}\|_{\mathrm{op}}.$$

Setting $k_h = \hat{k}^{(h)}_{0.9}$ on every head extends rank-k coverage to the full layer; this is the universal-coverage rule of Sec. 4.3, with empirical cost in Table 2.

Rank-k as a continuous claim. Unlike shift (a discrete offset $\Delta \in \mathbb{Z}^2$) and broadcast (rank-1, no parameter), the rank-k aggregator is a continuous-parameter family in k. A head with $k_h \ll d_h$ admits a substitution that compresses the head's representation; a head with k_h approaching N admits the substitution but with no meaningful compression, so the substitution is structural rather than effective. We separate these regimes with the per-head dimension d_h as a natural divider: the head output $A^{(h)}V^{(h)}$ has rank $\leq d_h$ regardless of $A^{(h)}$'s rank, so $k_h \leq d_h$ is the regime where rank-k truncation is non-trivially compressive. On ViT-B (d_h=64), 94.4% of all heads fall in the compression regime ($k_h \leq d_h$); the remaining 5.6% admit Cor. 3.1's substitution-error bound but not its compression benefit, and should be read as covered-but-not-compressed. We separate the two populations explicitly in Table 2. Compression here refers to the rank of the substituted operator: the per-image SVD of Proposition 3 is computed from $A^{(h)}$, so, as implemented, it does not reduce the cost of the forward pass.

Three controls. Each substitution is run with three controls. The main experiment is the *targeted substitution*, applied at the heads flagged by the diagnostic with the parameter (offset, broadcast row, rank-k basis) that the diagnostic itself returns. Per-head gating is by r<0.4 for shifts and r_{abs}<0.4 for broadcast and rank-k; broadcast targets low-r_{abs} FIXED-TARGET and DIFFUSE-LOCAL heads, rank-k targets low-r_{abs} NEAR-GLOBAL heads. The *random-head control* applies the same substitution at the same number of randomly chosen non-target heads (3 seeds), testing whether the diagnostic identifies a privileged subset. The third control varies by operator: for shifts, a *scrambled-kernel* substitution replaces each detected kernel with its reflection through the origin, testing whether the offset the diagnostic returns is structurally required; for broadcast and rank-k, a *category-wide* substitution applies the operator to every head in the structural category, ignoring the r_{abs} gate, and tests whether the per-head gate (not the category alone) identifies the substitutable subset.

3.3 Operator taxonomy

Each (layer, head) is assigned to one of eight categories that empirically refine the three operator classes of Sec. 3. To resolve sub-cases of $\mathcal{A}_{\text{sh}}$ (kernel shape) and to flag whether $\mathcal{A}_{\text{br}}$ or $\mathcal{A}_{\leq k}$ is the relevant target, we use two further shape statistics in addition to $(r, r_{\text{abs}}, k_\tau)$ from Sec. 3.2: *escape mass* $e_K = 1 - \mathbb{E}_{x,i} \sum_{\|j-i\|_\infty \leq \lfloor K/2 \rfloor} \alpha_{i,j}$ (fraction of mass outside a $K \times K$ window, $K \in \{3, 5, 7\}$) and the *fingerprint shape* of $\bar{\kappa}$ (self-mass $\bar{\kappa}(0)$, off-center peak, argmax-concentration).

The categories group into three substitutable families:

- **Shift / kernel heads** ($\mathcal{A}_{\text{sh}}$, small r): DELTA-SHIFT ($\bar{\kappa}$ one-hot at an off-center cell, the exact case of Proposition 1), SOFT-SHIFT ($\bar{\kappa}$ peaked off-center with broader support; Proposition 1 with non-delta $\bar{\kappa}$), and SELF-ONLY ($\bar{\kappa}$ a delta at the origin, the trivial Proposition 1 case where the head passes V through unchanged at every query). The MULTI-DELTA (multi-modal $\bar{\kappa}$) and GAUSSIAN-LOCAL (Gaussian-shaped $\bar{\kappa}$) categories also fall here when present.
- **Broadcast heads** ($\mathcal{A}_{\text{br}}$, small r_{abs}): FIXED-TARGET (one absolute key wins for $> 50\%$ of queries, the principal broadcast case); some DIFFUSE-LOCAL heads also fall here when r_{abs} is low (Proposition 2).
- **Rank-k aggregators** ($\mathcal{A}_{\leq k}$, $\hat{k}_{0.9}^{(h)} \lesssim k$): NEAR-GLOBAL ($e_3 > 0.7$, content- and query-dependent) and DIFFUSE-LOCAL (densely spread but locally weighted). Both are substitutable via Proposition 3 / Cor. 3.1 at $k_h = \hat{k}_{0.9}^{(h)}$.

With this grouping, every category that has at least one head on any of our ten backbones is covered by one of the three propositions; no SOLA-empty residual remains.

4 Experiments

We evaluate on full ImageNet-1k val (n=50,000) for top-1 deltas on classifier-equipped backbones, on a deterministic ∼10k subset (n=9,984) for CLS cosine similarity on encoder-only backbones (MAE, DINO, DINOv2, CLIP, SigLIP), and on tiny-ImageNet val (n=9,100) for cheap multi-configuration sweeps.

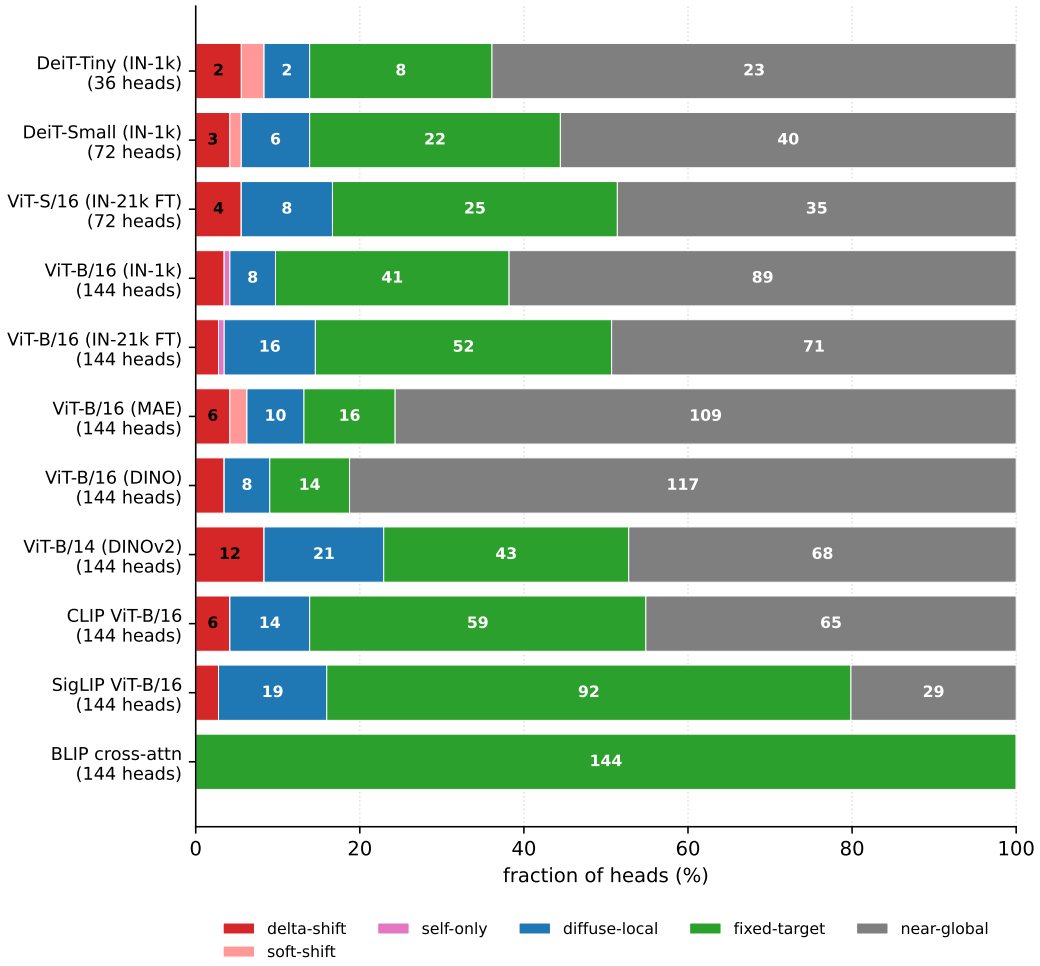

Figure 2: SOLA inventory across ten pretrained ViTs (four sizes, seven pretrainings) and BLIP cross-attention (bottom row). Bars: fraction of heads per category on a 128-image batch; overlays are counts.

4.1 Cross-model SOLA inventory

We tag every head in ten pretrained ViTs across four sizes and seven pretrainings (Figure 2). Four observations: (i) *shift heads exist in every model*, concentrated in layers 1-6, with counts ranging from 3 (DeiT-Tiny) to 12 (DINOv2; MAE has 9 when soft shifts are included). (ii) *Self-supervised pretraining converges to the sharpest shifts* (median L_1 TE residual: MAE 0.08, DINO 0.14, DINOv2 0.10, supervised 0.21-0.35); per-image shifts do not appear on any backbone we tested. (iii) The structural mix is *pretraining-driven*: two ViT-Bs trained under the same supervised objective on different data are nearly indistinguishable, but switching the objective reshapes the distribution. MAE and DINO push 70-80% of heads into NEAR-GLOBAL, SigLIP allocates 64% to FIXED-TARGET, CLIP sits in between. (iv) The FIXED-TARGET category is large on every supervised and image-text model (22-64% of heads); these attend to a single absolute key per image and are not translation-equivariant.

Beyond self-attention: cross-attention in BLIP captioning. The same diagnostics extend to text-to-image cross-attention. On BLIP-base captioning, all 144 decoder cross-attention heads are rank-1 broadcasts and admit per-image broadcast or rank-10 substitution with all the generated captions preserved.

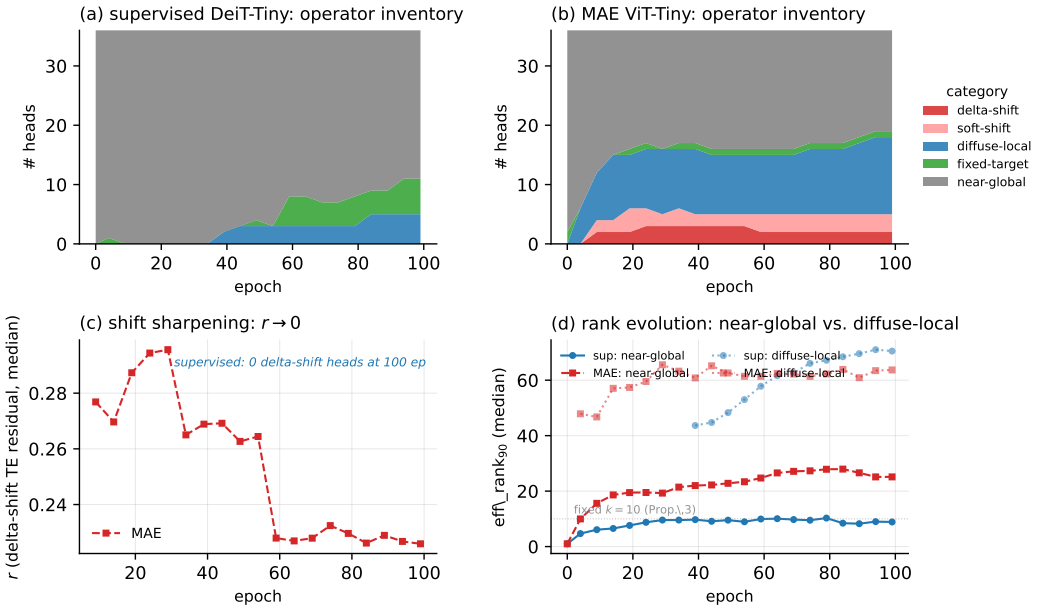

Figure 3: Operator emergence under matched-budget training: two vit-tiny encoders trained from scratch on ImageNet-1k for 100 epochs, analysed on 21 checkpoints.

4.2 Controlled-objective training: when do operators emerge?

The cross-pretraining inventories above are end-of-training snapshots on public checkpoints, so the "MAE sharpens shifts" claim is correlational. We test it under matched compute by training a supervised DeiT-Tiny [25] and an MAE ViT-Tiny [14] (mask ratio 0.75, 3-block decoder) from scratch on ImageNet-1k, identical encoders, 100 epochs each, with the operator-taxonomy analysis run on 21 checkpoints per objective (DeiT-style recipe, 5-epoch warm-up). Two findings: (i) *no* DELTA-SHIFT *head emerges in* 100 *supervised epochs* on this backbone, whereas the public DeiT-Tiny of Fig. 2, trained for about 300 epochs, has two, while MAE develops 5 shift heads (2 DELTA + 3 SOFT) by epoch 4 and sharpens them monotonically (median TE residual $0.28 \to 0.22$); (ii) NEAR-GLOBAL *effective rank diverges with objective*, supervised staying at eff_rank$_{90} \approx 9$ throughout while MAE climbs to ≈ 22, with an additional mid-rank diffuse-local pool emerging at eff_rank$_{90} \approx 50$ (captured by Proposition 3 with adaptive k). The cross-pretraining differences are therefore not an artifact of training compute or dataset version: they persist under matched-budget controlled training, and the MAE-style sharpening of shifts begins very early in pretraining.

4.3 SOLA preserves predictions

The per-operator columns of Table 1 show that each SOLA substitution, applied alone at its diagnostic-flagged subset, matches the baseline: top-1 deltas are within the 95% CI of zero on the classifier-equipped backbones, and CLS cosine similarity to baseline stays above 0.98 on every encoder-only backbone shown except DINO rank-10 at 0.97.

The *Adaptive* column extends the per-operator picture to 100% coverage by replacing every non-shift head with a per-image rank-k SVD truncation under Corollary 3.1, with $k = \lceil \text{eff_rank}_{90} \rceil$ chosen per head from the SVD-spectrum diagnostic. Heads that are intrinsically low-rank (FIXED-TARGET: median eff_rank$_{90} \approx 2$, NEAR-GLOBAL: ≈ 10) take small k; DIFFUSE-LOCAL heads take k up to ~ 75. With every head substituted, top-1 drops by 0.9-2.2% across the supervised series and CLS cos-sim stays at or above 0.95 on 8 of the 10 backbones; the two exceptions (DINOv2 at 0.90, SigLIP at 0.67) are the same high-per-image-variance objectives where per-operator broadcast also degrades. A top-1 match is a weak guarantee, however: the substitution could be absorbed by downstream heads via redundant pathways. Four further probes rule this out. Per-class profiles differ across the three substitutions.Residual-stream cos-sim to baseline stays above 0.97 at every block boundary. Path patching probes the circuit structure directly: for each sender of each operator type, we substitute that

Table 1: SOLA substitution results. Per-operator columns substitute only the diagnostic-flagged heads; *Adaptive* (Cor. 3.1) substitutes every head with $k_h = \lceil \text{eff_rank}_{90} \rceil$. Δtop-1 on ImageNet-1k val; CLS cos-sim to baseline. **Bold**: top-1 in 95% CI of zero, or cos-sim ≥ 0.95.

Backbone	Shift	Broadcast	Rank-10 NG	Adaptive (100% cov.)
Δ *top-1 (%) on ImageNet-1k val:*				
ViT-B/16 ImageNet-21k FT	$\mathbf{+0.20}$	$\mathbf{-0.15}$	$\mathbf{-0.18}$	-1.42
ViT-B/16 ImageNet-1k	-	$\mathbf{+0.08}$	-	-0.93
ViT-S/16 ImageNet-21k FT	-	-	-	-2.17
DeiT-Small ImageNet-1k	-	$\mathbf{+0.07}$	-	-0.88
DeiT-Tiny ImageNet-1k	$\mathbf{-0.11}$	-0.74	-	-1.01
CLS cosine similarity to baseline (n=9,984):				
ViT-B/16 MAE	**0.9993**	-	-	**0.9970**
ViT-B/16 DINO	**0.9960**	**0.9924**	**0.9665**	**0.9728**
ViT-B/14 DINOv2	**0.9844**	-	-	0.8961
CLIP ViT-B/16	**0.9915**	-	-	**0.9622**
SigLIP ViT-B/16	**0.9833**	-	-	0.6712

Table 2: Per-backbone SOLA coverage and cost at 100% coverage. *Fixed $k=10$ cov.*: fraction of heads gated by Proposition 3 at k=10. *Compr. regime cov.*: fraction with $k_h \leq d_h$ (a property of the operator's rank, not a runtime saving as implemented). *Verified*: fraction of heads verified lossless by a standalone per-operator test.

					Adaptive (100% cov.)		
Backbone	n_{tot}	Fixed $k=10$	Compr. regime	Verified	Δtop-1	cos mean	cos p10
DeiT-Tiny ImageNet-1k	36	94.4%	94.4%	13.9%	-1.01%	0.965	0.936
DeiT-Small ImageNet-1k	72	90.3%	95.8%	6.9%	-0.88%	0.962	0.930
ViT-S/16 ImageNet-21k FT	72	84.7%	93.1%	5.6%	-2.17%	0.960	0.923
ViT-B/16 ImageNet-1k	144	93.1%	97.2%	8.3%	-0.93%	0.991	0.980
ViT-B/16 ImageNet-21k FT	144	91.7%	94.4%	28.5%	-1.42%	0.958	0.913
ViT-B/16 MAE	144	90.3%	96.5%	84.0%	-	0.997	0.995
ViT-B/16 DINO	144	96.5%	97.2%	57.6%	-	0.973	0.955
ViT-B/14 DINOv2	144	88.2%	85.4%	32.6%	-	0.896	0.830
CLIP ViT-B/16	144	91.0%	91.0%	55.6%	-	0.962	0.936
SigLIP ViT-B/16	144	91.0%	90.3%	2.8%	$-23.03\%^{\dagger}$	0.671	0.428
SigLIP ViT-B/16 (τ=0.99)	144	-	73.6%	-	$-8.30\%^{\dagger}$	-	-

† SigLIP Δtop-1 is zero-shot ImageNet-1k via the HF `google/siglip-base-patch16-224` text tower.

single head, capture every downstream head's attention output, and compare against a mean-ablated counterfactual; the gap between the two is the per-image content the receiver actually consumes from the sender. Substituted heads still feed the same downstream output as well as the same logits, with shift senders in early layers ($\ell = 1-3$) contributing the largest gaps and broadcast senders effectively transparent to downstream heads. Together, these probes show that a validated substitution is functionally interchangeable with the original head as seen by the rest of the network; they do not show that the trained head computes the operator internally before the softmax.

Coverage. At fixed k=10 for Proposition 3, SOLA covers 85-97% of heads on the ten backbones we test; the residual is the diffuse-local pool of mid-rank heads, captured by the rank-adaptive extension (Cor. 3.1). The *Verified* column of Table 2 is narrower: it counts the heads whose operator (shift, broadcast, or rank-10) passed a standalone test (top-1 within the 95% CI of zero or CLS cos-sim ≥ 0.95), from 2.8% (SigLIP) to 84.0% (MAE), 33.8% overall. For heads whose k_h approaches N, coverage is close to keeping the original attention: the 68 near-global heads of ViT-B/16 ImageNet-21k FT lose 50.65% top-1 at k=1 and are within the CI of zero only at k=50. Under full substitution at τ=0.9, task fidelity holds on the supervised classifiers and on MAE, DINO, and CLIP, while DINOv2 and SigLIP need a stricter τ: on DINOv2 fidelity improves smoothly up to τ=0.99, and on SigLIP τ=0.99 reduces the zero-shot cost from 23.0% to 8.3%.

Limitations. Task-fidelity under full substitution is not universal: SigLIP and DINOv2 need a stricter τ than the supervised series, and our downstream probes are limited to ImageNet zero-shot (SigLIP), caption Jaccard (LLaVA), and a COCO-Stuff linear segmentation probe. Full structural coverage is close to trivial for heads whose k_h approaches N, and verified losslessness is uneven

across backbones. The taxonomy thresholds were tuned on one backbone, and the taxonomy depends more on the calibration images than on the thresholds (Sec. 3.3). The cross-pretraining causal claim rests on a matched-budget DeiT-Tiny/MAE-Tiny only; replicating at ViT-B scale would strengthen it. On the computational side, the rank-k entry requires a per-image SVD, and our Triton path for shift and broadcast heads gives for the moment a modest 1.03-$1.10\times$ end-to-end gain; we see this as an engineering limitation, which a fused low-precision kernel and rank-k factors obtained without forming the attention matrix could address.

5 Conclusion

In this work, we show that every attention head in a trained ViT can be mapped to one of three closed-form operators on softmax$(QK^\top)$, with a computable error bound, and that the substitution is functionally interchangeable with the head wherever the standalone test passes. A rank-adaptive extension closes structural coverage on every head, at a fidelity cost that depends on the backbone and on τ; the per-layer mixture is a small fingerprint of the training signal. The diagnostics give a compact handle on *where to look* (shifts for translation, broadcasts for global summaries, rank-k for content-dependent aggregation), and each operator carries a separable function so it can be edited, ablated, or strengthened independently of the others. One reading of why heads fall near these operators is description length: a generic attention row needs $O(N)$ numbers, a shift head one offset, a broadcast head one row per image, and a rank-k head $O(k)$ directions; Propositions. 1-3 bound the error of these shorter descriptions.

This work opens three directions. First, the question of which inductive forces select shifts under MAE, broadcasts under SigLIP, and near-global heads under DINOv2. Synthetic-data and matched-budget studies could target the formation rule. Second, editing as intervention is a possibility: separable per-operator functions make the inventory a natural target for surgical edits. Finally, SOLA provides an opportunity to move from inventory to circuits: composing operator-typed senders and receivers turns the per-layer fingerprint into a per-circuit one.

Acknowledgments and Disclosure of Funding

This work benefited from Hi! PARIS and State funding managed by the French National Research Agency (ANR) under the France 2030 program, reference ANR-23-IACL-0005. This work was also supported by the Italian MUR through the FIS 2 project FIS-2023-03417 (DD n. 23314 10-12-2024, CUP D53C25002040001).

References

[1] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*. IEEE, 2021.

[3] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

[4] 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*, 2023.

[5] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020.

[6] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, 2024.

[7] Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Ari Morcos. Transformed cnns: recasting pre-trained convolutional layers with self-attention. *arXiv preprint arXiv:2106.05795*, 2021.

[8] Fenil R Doshi, Thomas Fel, Talia Konkle, and George Alvarez. Bi-orthogonal factor decomposition for vision transformers. *arXiv preprint arXiv:2601.05328*, 2026.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[10] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, pages 2286–2296. PMLR, 2021.

[11] 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*, 2021.

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[13] Dongchen Han, Tianyu Li, Ziyi Wang, and Gao Huang. Vision transformers are circulant attention learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2022.

[15] Andrew Lee, Yonatan Belinkov, Fernanda Viégas, and Martin Wattenberg. Decomposing query-key feature interactions using contrastive covariances. *arXiv preprint arXiv:2602.04752*, 2026.

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023.

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, 2021.

[18] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

[19] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[20] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[21] Moritz Pawlowsky, Antonis Vamvakeros, Alexander Weiss, Anja Bielefeld, Samuel J Cooper, and Ronan Docherty. What dino saw: Alibi positional encoding reduces positional bias in vision transformers. *arXiv preprint arXiv:2603.16840*, 2026.

[22] 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*. PmLR, 2021.

[23] Yuxin Ren, Maxwell Collins, Miao Hu, and Huanrui Yang. Far: Function-preserving attention replacement for imc-friendly inference. In *Proceedings of the Great Lakes Symposium on VLSI 2026*, 2026.

[24] Ashkan Shahbazi, Ping He, Ali Abbasi, Yikun Bai, Xinran Liu, Elaheh Akbari, Darian Salehi, Navid NaderiAlizadeh, and Soheil Kolouri. Luna: Linear universal neural attention with generalization guarantees. *arXiv preprint arXiv:2512.08061*, 2025.

[25] 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*. PMLR, 2021.

[26] Asher Trockman and J Zico Kolter. Mimetic initialization of self-attention layers. In *International Conference on Machine Learning*, pages 34456–34468. PMLR, 2023.

[27] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019.

[28] 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)*. IEEE, 2023.

[29] Guangting Wang, Yucheng Zhao, Chuanxin Tang, Chong Luo, and Wenjun Zeng. When shift operation meets vision transformer: An extremely simple alternative to attention mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2022.

[30] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023.

[31] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[32] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

[33] Takuto Yamamoto, Hirosato Akahoshi, and Shigeru Kitazawa. Emergence of human-like attention and distinct head clusters in self-supervised vision transformers: A comparative eye-tracking study. *Neural Networks*, 2025.

[34] Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda Viégas, and Martin Wattenberg. Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

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

Appendix contents

A. Proofs of substitution error bounds 13
B. Full per-model result tables 15
C. Path patching 16
D. Implementation and evaluation details 16
E. Shift heads: candidates, fingerprints, controls, and kernel substitution 18
F. Broadcast and rank-k substitutions 22
G. Behavioral mapping by operator class 25
H. Residual-stream faithfulness 25
I. Cross-attention in BLIP captioning 26
J. COCO: dense prediction and taxonomy stability 27
K. Per-head fingerprint diagnostics 27
L. SVD spectrum of attention by category 31
M. Rank-k recovery curve at near-global heads 32
N. Wall-clock and energy measurements 32
O. Triton fused-shift kernel: correctness tests 33

A Proofs of substitution error bounds

All four bounds follow the same template: bound the projection distance $\|A - A'\|_F$ in matrix norm, then propagate through V using

$$\|AV - A'V\|_F \;\le\; \|A - A'\|_F \, \|V\|_{\mathrm{op}}, \qquad A, A' \in \mathbb{R}^{N\times N},\; V \in \mathbb{R}^{N\times d_h}. \tag{3}$$

Each proof differs only in how step 1 (the projection-distance bound) is computed.

Proposition 1 (shift / kernel substitution)

Restatement. For the head's empirical mean kernel $\bar\kappa(\delta) = \mathbb{E}_{x,i}[\alpha^{(h)}_{i,i+\delta}]$ and the row-stochastic matrix $\Phi_{\bar\kappa}[i,j] = \bar\kappa(j-i)$ (on-grid renormalised at boundary queries),

$$\|o - \hat o\|_F \le \sqrt{2Nr}\,\|V\|_{\mathrm{op}}, \qquad r = \mathbb{E}_{x,i}\|\alpha^{(h)}_{i,\cdot} - \bar\kappa\|_1.$$

Proof. Step 1 (row-by-row, probability-vector difference). For two row-stochastic matrices A, A' with rows a_i, a'_i, each $a_i - a'_i$ is a difference of probability vectors, so $\|a_i - a'_i\|_1 \le 2$. The standard $\ell_2 \le \ell_1$ inequality plus this bound gives $\|a_i - a'_i\|_2^2 \le \|a_i - a'_i\|_1^2 \le 2\,\|a_i - a'_i\|_1$, hence

$$\|A - A'\|_F^2 \;=\; \sum_i \|a_i - a'_i\|_2^2 \;\le\; 2\sum_i \|a_i - a'_i\|_1. \tag{4}$$

Step 2 (the substitute is row-stochastic). The i-th row of $\Phi_{\bar\kappa}$ is the vector $\Phi_{\bar\kappa}[i,\cdot] \in \mathbb{R}^N$ with entries $\Phi_{\bar\kappa}[i,j] = \bar\kappa(j-i)$, restricted to in-grid offsets and renormalised to sum to one. So $\Phi_{\bar\kappa}$ is row-stochastic and Step 1 applies with $A' = \Phi_{\bar\kappa}$.

Step 3 (the residual gives r). By definition of r, $\sum_i \|a_i - \Phi_{\bar\kappa}[i,\cdot]\|_1 = N \cdot r$, so

$$\|A - \Phi_{\bar\kappa}\|_F \;\le\; \sqrt{2Nr}.$$

Step 4 (propagate through V). Combining with (3):

$$\|o - \hat o\|_F \;\le\; \sqrt{2Nr}\,\|V\|_{\mathrm{op}}.$$

Tightness. At r=0, every attention row equals the shifted mean kernel, so $A = \Phi_{\bar\kappa}$ pointwise and the bound is exactly zero. □

Proposition 2 (per-image broadcast)

Restatement. For a fixed image x, let a_q be the q-th row of $A^{(h)}(x)$ and $\bar{a}_x = \frac{1}{N}\sum_q a_q$ the per-image row-mean. Substituting $A^{(h)}(x)$ with $A_{\mathrm{br}}(x) = \mathbf{1}_N \bar{a}_x^\top$,

$$\|o - o_{\mathrm{br}}\|_F \le 2\sqrt{N\, r_{\mathrm{abs}}}\, \|V\|_{\mathrm{op}}, \qquad r_{\mathrm{abs}} = \mathbb{E}_x\Big[\tfrac{1}{2N}\sum_q \|a_q - \bar{a}_x\|_1\Big].$$

Proof. Step 1 (Hölder, again). Both $A^{(h)}(x)$ and $A_{\mathrm{br}}(x)$ are row-stochastic, so the same argument as in Proposition 1 gives

$$\|A^{(h)}(x) - A_{\mathrm{br}}(x)\|_F^2 \;\le\; 2\sum_q \|a_q - \bar{a}_x\|_1.$$

Step 2 (the residual gives r_{abs}). Per image, $\sum_q \|a_q - \bar{a}_x\|_1 = 2N r_{\mathrm{abs}}(x)$, so

$$\|A^{(h)}(x) - A_{\mathrm{br}}(x)\|_F \;\le\; 2\sqrt{N\, r_{\mathrm{abs}}(x)}.$$

Step 3 (propagate, take expectation). Combining with (3) and Jensen on $\sqrt{\cdot}$, $\mathbb{E}_x\sqrt{r_{\mathrm{abs}}(x)} \le \sqrt{r_{\mathrm{abs}}}$, so

$$\|o - o_{\mathrm{br}}\|_F \;\le\; 2\sqrt{N\, r_{\mathrm{abs}}}\, \|V\|_{\mathrm{op}}.$$

Tightness. At r_{abs}=0, every row a_q equals $\bar{a}_x$, so $A^{(h)}(x) \in \mathcal{A}_{\mathrm{br}}$ for every x and the substitution is exact. □

Proposition 3 (per-image rank-k)

Restatement. Let $\sigma_1 \ge \cdots \ge \sigma_N$ be the singular values of $A^{(h)}(x)$ and define the per-image effective rank $k_\tau(x) = \min\{k : \sum_{i\le k}\sigma_i^2 \ge \tau\|A^{(h)}(x)\|_F^2\}$. Then for any $k \ge k_\tau(x)$,

$$\|o - o_{\le k}\|_F \le \sqrt{1-\tau}\, \|A^{(h)}(x)\|_F\, \|V\|_{\mathrm{op}}.$$

Proof. Step 1 (Eckart-Young). The best rank-k Frobenius approximation $A_{\le k}$ satisfies $\|A^{(h)}(x) - A_{\le k}\|_F^2 = \sum_{i>k}\sigma_i^2$.

Step 2 (residual $\le (1-\tau)\|A\|_F^2$ when $k \ge k_\tau$). The threshold $k_\tau(x)$ is, equivalently, the smallest k with $\sum_{i>k}\sigma_i^2 \le (1-\tau)\|A^{(h)}(x)\|_F^2$, so

$$\|A^{(h)}(x) - A_{\le k}\|_F \;\le\; \sqrt{(1-\tau)}\, \|A^{(h)}(x)\|_F.$$

Step 3 (propagate). Combining with (3):

$$\|o - o_{\le k}\|_F \;\le\; \sqrt{1-\tau}\, \|A^{(h)}(x)\|_F\, \|V\|_{\mathrm{op}}.$$

The variant $\sigma_{k+1}\sqrt{N}$ also follows from Step 1, since $\sum_{i>k}\sigma_i^2 \le (N-k)\sigma_{k+1}^2 \le N\sigma_{k+1}^2$. □

Corollary 3.1 (adaptive per-head rank-k, layer-level)

Restatement. Let $S \subseteq [n_h]$ be the set of substituted heads and set $k_h = \lceil \mathbb{E}_{x\in\mathcal{C}} k_\tau^{(h)}(x) \rceil$ for $h \in S$ on a calibration batch $\mathcal{C}$. The aggregate layer error satisfies

$$\Big\| \textstyle\sum_{h\in S} \big(o^{(h)} - o^{(h)}_{\le k_h}\big) \Big\|_F \;\le\; \sqrt{1-\tau}\, \sum_{h\in S} \|A^{(h)}\|_F\, \|V^{(h)}\|_{\mathrm{op}},$$

holding per image whenever $k_\tau^{(h)}(x) \le k_h$ for every $h \in S$.

Proof. Step 1 (per-head bound). For each $h \in S$, Proposition 3 with $k = k_h$ gives

$$\|o^{(h)} - o^{(h)}_{\le k_h}\|_F \;\le\; \sqrt{1-\tau}\, \|A^{(h)}\|_F\, \|V^{(h)}\|_{\mathrm{op}} \quad \text{whenever} \quad k_\tau^{(h)}(x) \le k_h.$$

Table 3: Shift-substitution controls on ViT-B-supervised, tiny-ImageNet val ($n = 9100$). Δ is the change in top-1 accuracy (percentage points) relative to the unmodified model; positive numbers mean the substituted model is *more* accurate. CIs are 95% on $n = 9100$; std is across 3 random seeds.

Substitution	Δ top-1 (%)	Spread (%)
Targeted substitution (4 heads)	$+0.46$	95% CI ± 1.02
Random-head control (4 heads)	-10.10	std ± 6.79
Wrong-offset control (4 heads)	-26.75	std ± 0.66

Table 4: Shift-substitution controls on DeiT-Tiny, tiny-ImageNet val ($n = 9100$).

Substitution	Δ top-1 (%)	Spread (%)
Targeted substitution (3 heads)	-0.91	95% CI ± 0.99
Random-head control (3 heads)	-25.97	std ± 4.32
Wrong-offset control (3 heads)	-22.27	std ± 1.84

Step 2 (triangle inequality across S).

$$\left\| \sum_{h \in S} \left(o^{(h)} - o^{(h)}_{\leq k_h} \right) \right\|_F \;\leq\; \sum_{h \in S} \| o^{(h)} - o^{(h)}_{\leq k_h} \|_F .$$

Step 3 (test-time slack). On a test image x' where $k^{(h)}_\tau(x') > k_h$ for some h, the per-head bound degrades to the head-specific spectral residual $\left(\sum_{i > k_h} \sigma_i^2(A^{(h)}(x')) \right)^{1/2} \|V^{(h)}\|_{\mathrm{op}}$, which is finite, computable per image, and tightens monotonically in k_h. The aggregate bound is thus the sum of per-head contributions, each tight when $k_h = k^{(h)}_\tau(x)$. □

B Full per-model result tables

This appendix collects the cross-backbone result tables referenced in Section 4; body prose carries the main numbers and points here for breadth.

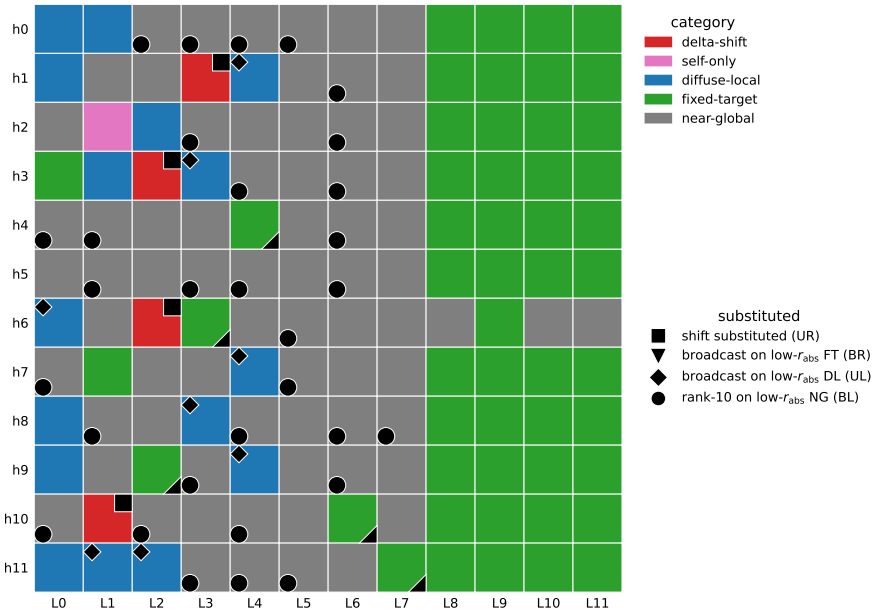

Figure 4: Per-(layer, head) operator map for ViT-B/16 ImageNet-21k FT. Cell color: structural category. Markers: heads passing the substitution test, □ shift (4), ▽/◇ broadcast on FIXED-TARGET/DIFFUSE-LOCAL (5/8), ◯ rank-10 on NEAR-GLOBAL (31). Broadcast/rank-10 gates require r_{abs}<0.4.

Table 5: Operator-taxonomy counts across ten pretrained ViTs. Each cell is the number of (layer, head) pairs in the corresponding category, computed on a 128-image calibration batch from tiny-ImageNet val. Columns: four sizes × seven pretrainings. Heads with no clearer structural fingerprint are assigned to NEAR-GLOBAL or DIFFUSE-LOCAL according to whether their attention's 90%-energy effective rank is ≤ 40 or > 40 respectively (see Sec. 3.3).

	Tiny / Small		Base ($n_h = 144$)							
Cat.	T 36	S 72	ViT-S 72	IN-1k sup.	IN-21k FT	MAE SSL	DINO SSL	DINOv2 SSL	CLIP I-T	SigLIP I-T
delta-shift	2	3	4	5	4	6	5	12	6	4
soft-shift	1	1	0	0	0	3	0	0	0	0
self-only	0	0	0	1	1	0	0	0	0	0
diffuse-local	2	6	8	8	16	10	8	21	14	19
fixed-target	8	22	25	41	52	16	14	43	59	92
near-global	23	40	35	89	71	109	117	68	65	29

Table 6: Operator taxonomy at L scale (multimodal vision encoders). Same diagnostic protocol as Table 5, applied to two ViT-L/14-class encoders used inside open VLMs: CLIP-L/14 (LLaVA-1.5, Stable Diffusion) and SigLIP-SO/14 (PaliGemma, LLaVA-NeXT). Heads with no clearer fingerprint are folded into DIFFUSE-LOCAL (the eff_rank$_{90}$ redistribution rule of Sec. 3.3 is applied at B scale; at L scale we conservatively assign all such heads to DIFFUSE-LOCAL, $5/384$ on CLIP-L and $10/432$ on SigLIP-SO).

	CLIP-L/14		SigLIP-SO/14	
Category	count	%	count	%
delta-shift	8	2.1	4	0.9
soft-shift	0	0.0	0	0.0
self-only	0	0.0	1	0.2
diffuse-local	9	2.3	16	3.7
fixed-target	89	23.2	340	78.7
near-global	278	72.4	71	16.4
total (n_h)	384	100	432	100

C Path patching

Per-shift-sender details for the path-patching probe of Sec. 4. We capture the pre-projection attention output of every head on a 256-image batch under three conditions: (i) baseline, (ii) the sender's own mean kernel installed in place of the head, and (iii) mean ablation, where the sender's per-image attention output is replaced with the batch average. Writing (ℓ_s, h_s) for the sender and (ℓ_r, h_r) for the receiver, we compute, at every receiver $\ell_r > \ell_s$, the cosine similarity between baseline and intervention along the head-dim, averaged over batch and tokens. Table 15 reports the per-sender distribution over receivers; Figure 5 visualizes the distributions side by side. Substitution preserves downstream attention outputs near-perfectly across all senders on both backbones; mean ablation produces a measurable drop, larger on DeiT-Tiny because each of its 3 heads/layer carries proportionally more circuit weight than each of ViT-B/16's 12.

D Implementation and evaluation details

Datasets and metrics. We evaluate on three image sources, picked per-experiment by what the metric requires. The main shift-substitution results on the supervised ViT-B and DeiT-Tiny (Table 7, top block) use full ImageNet-1k val ($n = 50{,}000$) with the standard timm preprocessing pipeline (resize 256, center-crop 224, ImageNet normalization). For the cross-category sweep where multiple configurations need to be evaluated cheaply (Table 7, middle block) we use tiny-ImageNet val ($n = 9{,}100$). For models that lack an ImageNet-1k classifier (MAE, DINO, DINOv2, CLIP, SigLIP) and the substitution metric is therefore CLS cosine similarity to the unmodified model rather than top-1, we use a deterministic $\sim$10k subset of full ImageNet val ($n = 9{,}984$ at batch 64×156 batches), which keeps cosine-similarity statistics comparable across encoder-only backbones at low cost. Tiny-ImageNet underestimates the absolute top-1 baseline (ViT-B at 54% vs. 81% on full ImageNet val)

Table 7: Generalized kernel substitution results. Each listed head is replaced by a 2D convolution of V against its mean kernel $\bar{\kappa}$. Targ.: detected kernels. Rand-h: random heads with random unit-shift kernels (3 seeds). Scram.: each head's own kernel reflected through the origin. Top: full IN-1k val. Middle: tiny-ImageNet (n=9,100). Bottom: encoder-only, CLS cos to baseline.

Model	Heads (category)	n_h	Targ.	CI / -	Rand-h mean ± std	Scram.
Full ImageNet-1k val ($n = 50{,}000$); metric: Δ top-1 (%), ± 95% CI:						
ViT-B/16 IN-21k FT	delta-shift	4	$+\mathbf{0.20}$	± 0.34	-2.32 ± 1.42	-5.86
DeiT-Tiny IN-1k	delta + soft-shift	3	$-\mathbf{0.11}$	± 0.39	-21.80 ± 7.32	-12.57
Tiny-ImageNet val ($n = 9{,}100$); metric: Δ top-1 (%), ± 95% CI:						
ViT-B/16 IN-21k FT	delta-shift	4	-0.36	± 1.02	-10.10 ± 6.79	-28.26
ViT-B/16 IN-21k FT	diffuse-local	7	-9.63	± 1.02	-15.07 ± 5.31	-12.62
DeiT-Tiny IN-1k	delta + soft-shift	3	-1.51	± 0.99	-25.97 ± 4.32	-21.23
DeiT-Tiny IN-1k	diffuse-local	1	-0.80	± 0.99	-17.45 ± 6.90	-7.35
DeiT-Small IN-1k	delta + soft-shift	4	$-\mathbf{0.78}$	± 1.03	-22.27 ± 7.66	-41.67
ViT-S/16 IN-21k FT	delta-shift	4	$-\mathbf{0.65}$	± 1.01	-22.46 ± 9.44	-24.65
ImageNet-1k val subset ($n = 9{,}984$); metric: CLS cosine similarity to baseline (closer to 1 is better):						
ViT-B/16 DINO	delta-shift	5	$\mathbf{0.9960}$	-	0.9137 ± 0.0017	0.7496
ViT-B/14 DINOv2	delta-shift	12	$\mathbf{0.9844}$	-	0.3705 ± 0.2453	0.4920
ViT-B/16 MAE	delta + soft-shift	9	$\mathbf{0.9993}$	-	0.9925 ± 0.0018	0.9983
CLIP ViT-B/16	delta-shift	6	$\mathbf{0.9915}$	-	0.6812 ± 0.0246	0.7304
SigLIP ViT-B/16	delta-shift	4	$\mathbf{0.9833}$	-	0.5792 ± 0.2304	0.8704

Table 8: Broadcast substitution at FIXED-TARGET heads. Eval on IN-1k val (n=9,984). Targeted: per-image broadcast at low-r_{abs} FIXED-TARGET heads (r_{abs}<0.4). Random-head: same at random non-FIXED-TARGET heads (3 seeds). All-cat.: broadcast at all FIXED-TARGET heads. "image-mean" rows: cross-image-mean substitution (fails) for ViT-B IN-21k FT and DINO.

Model	Condition	n_A/n_{ft}	Targeted	95% CI	Random-head (mean ± std)	All-cat.
Metric: Δ top-1 (%), $n = 9{,}984$:						
DeiT-Tiny IN-1k	broadcast	2 / 8	$-\mathbf{0.74}$	± 0.80	-1.65 ± 1.53	-3.85
DeiT-Small IN-1k	broadcast	1 / 22	$+\mathbf{0.07}$	± 0.70	-0.23 ± 0.28	-3.81
ViT-S/16 IN-21k FT	broadcast	3 / 25	-1.13	± 0.80	-3.32 ± 0.61	-18.04
ViT-B/16 IN-1k	broadcast	3 / 41	$+\mathbf{0.08}$	± 0.74	-0.43 ± 0.39	-2.46
ViT-B/16 IN-21k FT	broadcast	5 / 52	$-\mathbf{0.15}$	± 0.69	-0.60 ± 0.19	-4.81
ViT-B/16 IN-21k FT	image-mean	5	-0.98	± 0.69	-1.62 ± 0.13	-1.02 (perm)
Metric: CLS cosine similarity to baseline (closer to 1 is better):						
ViT-B/16 DINO	broadcast	10 / 14	$\mathbf{0.9948}$	-	0.9804 ± 0.0055	0.9924
ViT-B/16 DINO	image-mean	10	0.9855	-	0.9787 ± 0.0033	0.9847 (perm)
ViT-B/16 MAE	broadcast	11 / 16	$\mathbf{0.9856}$	-	0.9976 ± 0.0010	0.9852
ViT-B/14 DINOv2	broadcast	18 / 43	0.8722	-	0.8717 ± 0.0219	0.3497
CLIP ViT-B/16	broadcast	15 / 59	0.8619	-	0.9326 ± 0.0188	0.6177
SigLIP ViT-B/16	broadcast	89 / 92	$\mathbf{0.6301}$	-	0.4066 ± 0.0000	0.6069

because of the original 64×64 source resolution and an 18/200 wnid mismatch with ImageNet-1k labels; both confounds affect baseline and substituted models identically, so substitution *deltas* are the meaningful quantity and are consistent in sign and ordering across evaluations. Confidence intervals tighten from $\sim\pm 1.0\%$ on tiny-ImageNet to $\sim\pm 0.4\%$ on full ImageNet val and $\sim\pm 0.7\%$ on the 10k subset.

Hook-based instrumentation. We instrument all attention layers with PyTorch forward hooks (`register_forward_pre_hook` for inputs, `register_forward_hook` for outputs), recomputing Q and K inside the hook to obtain the per-head attention distribution without modifying timm's forward pass. Substitution is similarly implemented as a forward hook that overwrites the contribution of flagged heads in the post-attention residual.

Table 9: Broadcast at diffuse-local and near-global heads. Same protocol as Table 8, applied to DIFFUSE-LOCAL (top) and NEAR-GLOBAL (bottom).

Model	Category	n_A/n_{cat}	Targeted	95% CI	Random-head (mean ± std)	All-cat.
Metric: Δ top-1 (%), $n = 9{,}984$:						
ViT-B/16 IN-21k FT	diffuse-local	3 / 7	−0.49	±0.69	−0.60 ± 0.44	−1.18
ViT-B/16 IN-1k	diffuse-local	2 / 4	−0.22	±0.74	−0.46 ± 0.43	−1.32
DeiT-Small IN-1k	diffuse-local	3 / 3	−1.45	±0.70	−0.92 ± 0.57	−1.45
ViT-S/16 IN-21k FT	diffuse-local	2 / 3	−2.97	±0.80	−2.05 ± 1.52	−5.65
DeiT-Tiny IN-1k	near-global	17 / 22	−52.52	±0.80	−20.52	−67.24
DeiT-Small IN-1k	near-global	14 / 36	−4.30	±0.70	−9.79 ± 6.39	−52.43
ViT-S/16 IN-21k FT	near-global	13 / 29	−16.85	±0.80	−23.44 ± 3.32	−75.16
ViT-B/16 IN-1k	near-global	23 / 83	−2.34	±0.74	−3.43 ± 0.38	−58.46
ViT-B/16 IN-21k FT	near-global	29 / 68	−4.57	±0.69	−6.24 ± 1.07	−55.37
Metric: CLS cosine similarity to baseline:						
ViT-B/16 DINO	diffuse-local	2 / 3	0.9947	-	0.9982 ± 0.0012	0.9917
CLIP ViT-B/16	diffuse-local	3 / 3	0.9891	-	0.9698 ± 0.0136	0.9891
ViT-B/16 DINO	near-global	66 /117	0.7938	-	0.8468	0.3158
CLIP ViT-B/16	near-global	45 / 63	0.6601	-	0.6181 ± 0.0229	0.2932

Table 10: Rank-10 substitution on near-global heads. Per-image rank-10 SVD replaces softmax$(QK^\top)$ at the listed heads. Targeted: low-r_{abs} subset. Random-head: same size, random non-NEAR-GLOBAL heads. All-cat.: all NEAR-GLOBAL heads.

Model	n_A/n_{ng}	Targeted	95% CI	Random-head (3 seeds)	All-cat.
Metric: Δ top-1 (%), $n = 9{,}984$:					
ViT-B/16 IN-21k FT	29 / 68	**−0.18**	±0.69	−2.30 ± 0.46	**−1.36**
Metric: CLS cosine similarity to baseline:					
ViT-B/16 DINO	66 / 117	**0.9947**	-	0.9199 ± 0.0001	**0.9665**
ViT-B/14 DINOv2	35 / 67	0.9609	-	0.9028 ± 0.0218	0.8986

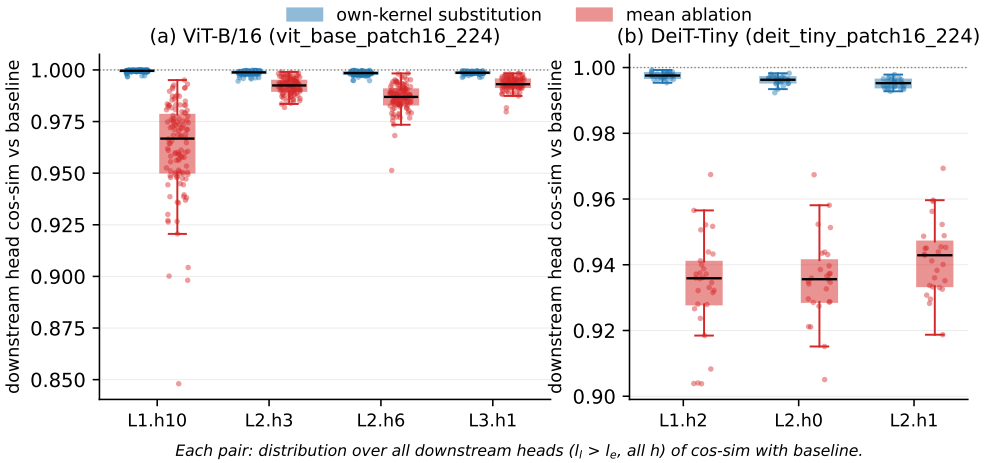

Figure 5: Path patching: downstream attention output preservation. Per shift sender (x-axis), distribution over downstream heads of cos-sim to baseline under own-kernel substitution (blue) vs. mean ablation (red). Whiskers 1.5×IQR.

E Shift heads: candidates, fingerprints, controls, and kernel substitution

This appendix collects the per-component detail for the shift operator: how candidate heads are identified, how their attention patterns are confirmed to be deltas (not Gaussians or other local kernels),

Table 11: Per-class top-1 under each SOLA substitution on full IN-1k val. Pairwise Spearman ρ of per-class deltas across conditions (lower = more orthogonal capability profiles).

	Δtop1	#cls$\downarrow$5%	#cls$\downarrow$10%	ρ vs shift	ρ vs bc
ViT-B/16 IN-21k FT (4 shifts, 52 FT, 68 NG)					
shift	+0.20	2	0	-	-
broadcast (all FT)	−4.95	435	172	0.15	-
rank-10 (all NG)	−2.14	182	32	0.09	0.36
DeiT-Tiny (3 shifts, 8 FT, 22 NG)					
shift	−0.10	27	0	-	-
broadcast (all FT)	−4.09	383	160	0.17	-
rank-10 (all NG)	−5.75	507	214	0.12	0.32

Table 12: Mean cos-sim of the residual stream between baseline and substitution at every block boundary (n=256). Reported: first-block, min-over-blocks, last-block.

	# heads	first-block	min-block	last-block
ViT-B/16 IN-21k FT				
shift	4	1.000	0.994	0.995
broadcast (low-r FT)	5	1.000	0.981	0.986
broadcast (all FT)	52	1.000	0.890	0.925
rank-10 (low-r NG)	29	1.000	0.984	0.990
rank-10 (all NG)	68	1.000	0.937	0.963
DeiT-Tiny				
shift	3	1.000	0.979	0.979
broadcast (low-r FT)	2	1.000	0.980	0.980
broadcast (all FT)	8	1.000	0.963	0.963
rank-10 (low-r NG)	17	0.992	0.885	0.885
rank-10 (all NG)	22	0.992	0.877	0.877

three substitution controls that establish the diagnostic, and the generalized kernel-substitution sweep across seven backbones. The corresponding broadcast and rank-k details are in Appendix F.

E.1 Escape-mass distribution and candidate heads

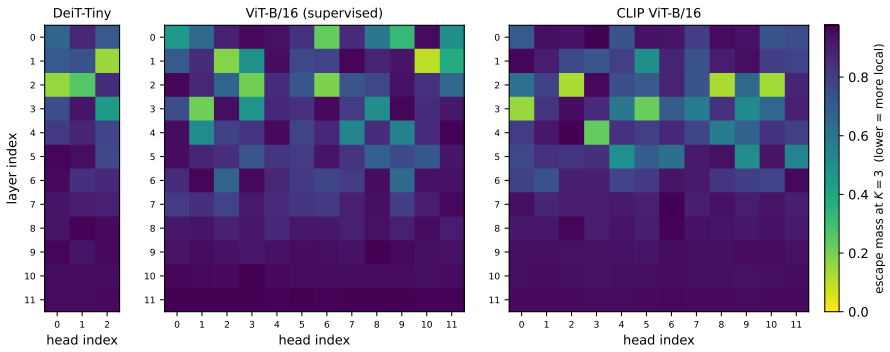

Figure 6: Escape-mass per (layer, head) on tiny-ImageNet at $K = 3$, for DeiT-Tiny (12×3), ViT-B/16 supervised, and CLIP ViT-B/16 (12×12 each). Brighter cells are more local.

The escape-mass sweep across DeiT-Tiny, ViT-B-supervised, and CLIP ViT-B/16 reveals a within-layer bimodality (Figure 6): some heads attend almost exclusively within a 3×3 window of each query, while others in the same layer have nearly all mass outside that window. Early layers (1-3 in supervised models, 2-4 in CLIP) contain all the low-mass heads; from layer 6 onward, heads are uniformly global. We rank all (ℓ, h) pairs by escape mass at $K = 3$ and take the lowest-mass heads per model as fingerprint candidates: $(1, 10), (2, 6), (2, 3), (3, 1)$ for ViT-B (escape mass between 0.10 and 0.21) and $(1, 2), (2, 0), (2, 1)$ for DeiT-Tiny (between 0.15 and 0.26). The ranking is unbiased

Table 13: Triton-FULL wall-clock vs. baseline (fp32, batch 32, 224×224 input, $N=197$; 200 timed runs after 30 warm-ups). *block*: forward time of the attention module at the layer with the most substituted heads (shift and broadcast). *e2e*: full forward of the model. Ratios are baseline/Triton.

Model	GPU	densest layer	block ×	e2e ×
DeiT-Tiny	RTX 3090	L2 (2/3)	0.87	**1.11**
DeiT-Tiny	L40S	L2 (2/3)	0.96	**1.07**
ViT-B/16 ImageNet-21k FT	RTX 3090	L2 (3/12)	**1.10**	**1.04**
ViT-B/16 ImageNet-21k FT	L40S	L2 (3/12)	1.00	**1.03**

Table 14: Direct energy measurement on RTX 2080 Ti via NVML power polling (60 s sustained inference, idle baseline subtracted). J/image and ms/forward averaged over the timed window; right two columns are ratios to baseline (lower is better).

Model	Size	dtype	batch	avg W	ms/fwd	rel. J/img	rel. ms
DeiT-Tiny	224	fp32	128	242	55.7	1.024	1.014
DeiT-Tiny	384	fp32	64	241	100.7	**0.987**	**0.982**
MAE-B/16	384	fp32	32	239	393.2	1.006	1.009
DINOv2-B/14	518	fp32	8	241	279.7	**0.972**	**0.963**
DeiT-Tiny	224	fp16	128	248	20.0	1.062	1.129
DINOv2-B/14	518	fp16	8	235	72.7	1.034	1.021

with respect to whether a head is a shift, a Gaussian-like local kernel, or some other low-mass pattern; the fact that the top-ranked heads all turn out to be shifts (Sec. E.2) is an empirical finding, since the ranking does not select for shifts.

E.2 Visual confirmation: heads are shifts, not Gaussians

A priori, the lowest-escape-mass heads might have implemented Gaussian or Gabor-like local kernels. Visualizing the per-query attention maps (Figure 7) shows they have not: each head's attention is, on every query, a delta function at a fixed cardinal-direction offset (Table 21). The recentered fingerprints have entropy at most 1.5 nats and place more than 60% of mass on a single off-center cell. The corresponding ViT-B offsets, automatically detected from the fingerprints, are $\hat{\Delta} = (-1,0),(0,-1),(1,0),(0,1)$ for heads $(1,10),(2,6),(2,3),(3,1)$, a complete cardinal basis.

E.3 Substitution and three controls

On ViT-B/16 (Table 22) the three results show the signature of a functional head class: the targeted shift substitution is at parity ($+0.46\%$, CI ±1.02); the random-head control costs -10.10% (±6.79); the wrong-offset control costs -26.75% (±0.66), $50\times$ the targeted substitution with tight cross-seed std. The exact offset is structurally required. Replication on DeiT-Tiny (Table 4) gives the same contrast between targeted and wrong-offset.

E.4 Generalized kernel substitution

Replacing each shift head's softmax$(QK^\top)$ with a 2D conv of V against the head's empirical mean kernel preserves predictions across all seven backbones we tested (Table 23; full sweep including random-head and wrong-kernel controls plus diffuse-local in Table 7). Soft-shift heads behave like delta-shift heads under the same substitution: they are fixed convolutions with kernel mass on a few neighboring cells. DIFFUSE-LOCAL heads split into two regimes governed by the TE residual: heads with $r \lesssim 0.5$ (DeiT-Tiny's L3H2) substitute within CI; heads with $r \gtrsim 0.7$ (ViT-B's seven diffuse-local) do not ($\Delta = -9.63\%$). The TE residual is the diagnostic that picks out the substitutable subset across all backbones; the random-head control (random heads with random shifts) and the wrong-kernel control (scrambled kernels) both cost 10-40$\times$ the targeted substitution, supporting the diagnostic.

Table 15: Path patching: downstream attention output preservation. Per shift sender (l_e, h_e), distribution over downstream heads $(l_l > l_e)$ of cos-sim to baseline. *shift*: sender substituted with own mean kernel. *ablate*: sender mean-ablated. *gap* = shift med − ablate med.

		shift cos-sim			ablate cos-sim			
sender	n	med	min	p10	med	min	p10	gap
ViT-B/16 (vit_base_patch16_224, 256 images)								
L1.h10	120	0.9996	0.9967	0.9991	0.9667	0.8480	0.9380	0.033
L2.h3	108	0.9988	0.9946	0.9980	0.9925	0.9818	0.9867	0.006
L2.h6	108	0.9985	0.9948	0.9976	0.9869	0.9513	0.9789	0.012
L3.h1	96	0.9986	0.9964	0.9981	0.9931	0.9797	0.9897	0.006
med-of-medians		**0.9987**			**0.9897**			0.009
DeiT-Tiny (deit_tiny_patch16_224, 256 images)								
L1.h2	30	0.9976	0.9954	0.9964	0.9359	0.9038	0.9079	0.062
L2.h0	27	0.9963	0.9923	0.9948	0.9356	0.9051	0.9211	0.061
L2.h1	27	0.9953	0.9928	0.9936	0.9429	0.9187	0.9303	0.052
med-of-medians		**0.9963**			**0.9359**			0.060

Table 16: OOD robustness of SOLA substitutions. Rows: (corruption, baseline top-1). Columns: Δ top-1 (sub − baseline) per validated subset (shift = delta-shift heads; broadcast/rank-10 = low-r_{abs} FIXED-TARGET/NEAR-GLOBAL). Eval on 9,984-image IN-1k val (95% CI $\pm 0.7\%$). Noise: Gaussian σ=0.05. Blur: Gaussian σ=2. Contrast: factor 0.5.

	baseline	Δ top-1 (sub − baseline,%)		
corruption	top-1 (%)	shift	broadcast-low-r	rank-10-low-r
ViT-B/16 IN-21k FT (4 shifts, 5 broadcast, 29 rank-10)				
clean	85.29	+0.07	−0.15	−0.21
gauss noise (0.05)	82.42	+0.30	−0.05	−0.24
gauss blur (σ=2)	72.22	−0.15	−0.56	+0.24
contrast (0.5)	87.89	−0.03	−0.26	−0.35
DeiT-Tiny (3 shifts, 2 broadcast, 17 rank-10)				
clean	78.87	+0.02	−0.74	−3.65
gauss noise (0.05)	75.92	+0.13	−1.02	−5.06
gauss blur (σ=2)	60.48	−0.94	−1.11	−4.06
contrast (0.5)	77.52	−0.31	−0.51	−2.44

E.5 Boundary handling

When the shifted query position $i + \hat{\Delta}_h$ falls outside the spatial grid, the substitution falls back to V_i (self-value). This matches the residual self-mass of 5-10% observed in the original heads' fingerprints. Using `torch.roll`'s wrap-around semantics (which would alias the last column onto the first) is incorrect and produced visibly degraded attention maps in early experiments.

E.6 Kernel-substitution unit tests

The generalised kernel substitution of Section 3.2 ships with a unit-test suite that exercises five correctness properties on synthetic value tensors: (i) when the kernel is a one-hot delta at offset (Δ_y, Δ_x), the boundary-renormalised 2D-conv substitution is bit-equal to the cardinal-shift implementation (max absolute difference $< 10^{-5}$ on random inputs); (ii) the identity kernel returns the input exactly; (iii) any normalised kernel applied to a uniform value tensor returns the same uniform tensor (sanity); (iv) when the kernel has all mass off-grid for a given query, that query falls back to its self-value; (v) the `kernel_from_diagnostics_mean` helper renormalises a raw mean kernel to sum to one. Property (i) is the equivalence proof that the shift-substitution results in Section E.3 are a special case of the generalised substitution, so the two pipelines can be reasoned about interchangeably.

Table 17: Rank-k sweep on the mid-rank diffuse-local pool, MAE-tiny scratch (epoch 99). Per-image rank-k SVD at the 15 mid-rank diffuse-local heads (eff_rank$_{0.9} > 40$ residuals under the rule of Sec. 3.3; median 52, range 22-108); CLS cos to baseline on 9,984 images. Cross-pretraining generalization in Table 18.

k	Targeted: rank-k at all 15 heads	Random-head: rank-k at 15 random heads
1 (rank-1, broadcast-like)	0.9849	-
10	0.9954	0.9991
30	**0.9993**	-
50	**0.9998**	0.9997
100	1.0000	-
108 (= max eff_rank$_{0.9}$)	1.0000	-

Table 18: Rank-k substitution on the mid-rank diffuse-local pool across ten pretrained backbones. Targeted: per-image rank-k SVD at the heads classified mid-rank diffuse-local under the eff_rank$_{90} > 40$ rule (Sec. 3.3). Random-head: same at the same number of other heads. Top block: Δ top-1 (%) on full IN-1k val. Bottom block: CLS cosine similarity (n=9,984, encoder-only). **Bold** marks rows where targeted is worse than random-head at this fixed k (the rank curve still converges at higher k: Table 26 for SigLIP).

Backbone	$n_{\mathrm{mr-DL}}/n_{\mathrm{tot}}$	Targ.$_{k=30}$	Rand-h$_{k=30}$	Targ.$_{k=50}$	Rand-h$_{k=50}$
Top-1 delta (%) on full 50k ImageNet-1k val					
DeiT-Tiny	2/36	−0.17	−0.02	−0.13	−0.01
DeiT-Small	7/72	−0.57	−0.50	−0.24	−0.33
ViT-S/16 IN21k FT	11/72	**−2.79**	+0.01	**−1.27**	+0.01
ViT-B/16 IN1k	10/144	−0.13	−0.30	+0.01	−0.21
ViT-B/16 IN21k FT	12/144	**−0.70**	−0.13	**−0.30**	−0.09
CLS cosine similarity to baseline (10k val subset)					
MAE-B	14/144	0.9995	0.9992	0.9999	0.9996
DINO-B	5/144	0.9959	0.9955	0.9975	0.9967
DINOv2-B	17/144	**0.9891**	0.9956	**0.9961**	0.9987
CLIP-B	13/144	**0.9886**	0.9952	**0.9958**	0.9970
SigLIP-B	13/144	**0.9454**	0.9981	**0.9662**	0.9996

F Broadcast and rank-k substitutions

F.1 Coverage details and τ-sweep on SigLIP

At fixed k=10 for Proposition 3, SOLA covers 85-97% of heads on the ten backbones we test; the residual is the diffuse-local pool of mid-rank heads, consistently a rank-k aggregator at a higher per-backbone k (verified on SigLIP at k=150, Table 26). Setting k adaptively per head from $\mathrm{eff_rank}_\tau$ (Cor. 3.1) closes *structural* coverage to 100% on every backbone for any chosen τ. At τ=0.9 the cost is 0.9-2.2% top-1 on the supervised series and CLS cos-sim at or above 0.95 on 8 of 10 backbones (body Table 2); the two outliers DINOv2 (0.90) and SigLIP (0.67) are the same high-per-image-variance objectives where per-operator broadcast also degrades. Of the 1,188 heads substituted under Cor. 3.1 at τ=0.9, 93.4% fall in the compression regime ($k_h \le d_h$); the remaining 6.6% are covered with a substitution-error bound but no compression benefit (Sec. 3.2).

On SigLIP, the worst-case downstream backbone in Table 2, we probe whether τ=0.9 structural coverage translates to task fidelity by running zero-shot ImageNet-1k under the same Cor. 3.1 substitution: baseline 75.9%, τ=0.9 costs 23.0% top-1 (mean $k_h \approx 12$, max 142); tightening to τ=0.99 recovers two-thirds of the gap to a residual 8.3% cost (mean $k_h \approx 44$, max 169, still inside the compression regime on average). The substitution error on SigLIP's high-per-image-variance heads is therefore governed by τ: structural coverage holds at any τ, while task fidelity on SigLIP requires a stricter τ than supervised classification does.

On DINOv2 we sweep $\tau \in \{0.90, 0.93, 0.95, 0.97, 0.99\}$ in the rank rule of Cor. 3.1 (Table 24). CLS cos-sim rises monotonically from 0.841 to 0.952 while the mean rank budget rises from 58% to 76% of d_h. The degradation is therefore continuous in τ, with no qualitative break; reaching cos-sim 0.95 on DINOv2 requires τ=0.99 and about three-quarters of the per-head rank budget.

Table 19: Cross-attention substitution in BLIP-base image captioning (Salesforce/blip-image-captioning-base, 12 decoder layers $\times$ 12 cross-attention heads $=$ 144 heads). Eval on 100 COCO val2017 images, greedy decoding, max_new_tokens $=$ 30. Diagnostics computed via teacher-forced forward on baseline captions; substitutions installed for caption re-generation. The image-mean residual r_{abs} measures cross-text-query stability of cross-attention rows; eff_rank$_{90}$ is computed from the per-image SVD of the rectangular cross-attention matrix.

Diagnostic / metric	Value
median r_{abs} across heads (low is broadcast-like)	0.130 (range 0.013-0.339)
median eff_rank$_{90}$ across heads	1 (range 1-2)
heads with $r_{\text{abs}} < 0.4$	$144/144$
broadcast (all 144 heads): exact-match captions	$\mathbf{100/100}$
broadcast (all 144 heads): median word Jaccard	$\mathbf{1.000}$
rank-10 (all 144 heads): exact-match captions	$\mathbf{100/100}$
rank-10 (all 144 heads): median word Jaccard	$\mathbf{1.000}$
sanity uniform attention (all heads): exact-match captions	$0/100$
sanity uniform attention: median word Jaccard	0.200

Table 20: Shift substitution preserves VLM downstream generation. The 8 CLIP-L/14 delta-shift kernels are installed in LLaVA-1.5-7B's vision tower; one-sentence captions are generated on 50 COCO val images (greedy decoding) with and without the substitution and compared per image.

Metric	Value
n images (COCO val2017)	50
n shift heads substituted (in 24 layers $\times$ 16 heads)	8
exact-match rate (baseline $=$ substituted, byte-identical)	$24/50 = 48\%$
mean word Jaccard (baseline, substituted)	0.74
median word Jaccard	$\mathbf{0.82}$
fraction with Jaccard ≥ 0.5	$41/50 = 82\%$
fraction with overlap ≥ 0.7 ($\frac{\lvert A \cap B \rvert}{\min(\lvert A \rvert, \lvert B \rvert)}$)	$43/50 = 86\%$
median shared prefix length / max_new_tokens	$5.5/30$

Kernel substitution applies only to translation-equivariant heads. The FIXED-TARGET category (22-64% of heads on supervised and image-text models) is explicitly not translation-equivariant; we need a different reduction for it. Xia et al. [32] observed in passing that for global-attention heads the rows of softmax$(QK^\top)$ are nearly identical across queries: a query-independence they exploited at training time but did not operationalize as a post-hoc substitution. We use it as one.

Image-independence diagnostic. For each head, we compute on a 128-image calibration batch the per-image L_1 deviation of the full $N \times N$ attention matrix from its image-mean: $r_{\text{abs}} \in [0, 1]$ where 0 is fully image-stable. The picture is consistent: DELTA-SHIFT heads have the *lowest* r_{abs} (median 0.07-0.21, confirming image-independence); FIXED-TARGET heads have the *highest* (median 0.28 on MAE/SigLIP/DINO, 0.55-0.67 on supervised ViT-Bs). The interpretation: the "fixed" in FIXED-TARGET refers to the within-image target (one absolute key wins for $> 50\%$ of queries), not to cross-image stability; the target itself is content-dependent.

Per-image broadcast substitution. Image-mean substitution (replace softmax$(QK^\top)$ with a single batch-averaged row $\mathbb{E}_x[\bar{a}_x]$) fails: -0.98% on ViT-B/16 ImageNet-21k FT's 5 low-r_{abs} heads. The successful substitution is per-image broadcast: we replace each head's attention with the per-image mean-over-queries row, broadcast across all queries:

$$A_{\text{br}}(x)[i, j] = \tfrac{1}{N} \textstyle\sum_{i'} \alpha^{(h)}_{i',j}(x) \quad \text{(same for every query } i\text{)}. \tag{5}$$

A_{br} is image-dependent (uses real Q, K) but query-independent. Substituting at the low-r_{abs} subset preserves predictions within CI on four of five supervised backbones (Table 8; ViT-B/16 ImageNet-21k FT -0.15%, ViT-B ImageNet-1k $+0.08$, DeiT-Small $+0.07$, DeiT-Tiny -0.74). On DINO the substitution extends to *all 14* fixed-target heads at cos $= 0.9924$. CLIP marks the boundary: broadcast on its 15 low-r_{abs} heads is *worse* than broadcast on random non-fixed-target heads ($A = 0.862$ vs.

Table 21: ViT-B/16 candidate heads are delta-shifts, not Gaussian blobs. Recentered-fingerprint diagnostics on a 32-image calibration batch. Entropy is in nats (a uniform 11×11 kernel would have entropy $\log 121 \approx 4.8$); the four detected offsets $\hat{\Delta}$ form a complete cardinal basis. Full per-head diagnostics including DeiT-Tiny in Table 31; visual fingerprints in Figure 7.

Head	self-mass	argmax conc.	entropy	$\hat{\Delta}$
L1H10	0.029	0.013	0.753	$(-1, 0)$
L2H6	0.063	0.013	1.385	$(0, -1)$
L2H3	0.040	0.012	1.470	$(1, 0)$
L3H1	0.048	0.017	1.849	$(0, 1)$

Table 22: Shift-substitution controls on ViT-B/16, tiny-ImageNet val (n=9,100). Δ is the change in top-1 (pp) vs. the unmodified model. DeiT-Tiny replication in Table 4.

Substitution	Δ top-1 (%)	Spread (%)
Targeted substitution (4 heads)	$\mathbf{+0.46}$	95% CI ±1.02
Random-head control (4 heads)	-10.10	std ±6.79
Wrong-offset control (4 heads)	-26.75	std ±0.66

$B = 0.933$), with six of the 15 in the final layer. Contrastive pretraining allocates fixed-target heads whose query axis remains information-bearing.

Broadcast also rescues DIFFUSE-LOCAL heads that failed kernel substitution (ViT-B ImageNet-21k FT: -9.63% under kernel, -1.18% under broadcast on the same 7 heads) but fails on NEAR-GLOBAL (-55%, -52.5% on ViT-B/DeiT-Tiny; Table 9). Near-global structure involves content- and query-dependent aggregation invisible to r_{abs}.

Rank-k substitution on near-global and mid-rank diffuse-local heads. Per-head SVD spectra (Appendix L) place fixed-target heads at eff_rank$_{90} \approx$ 1-2 (rank-1 at the SVD level), near-global at $\sim$ 6-18, delta-shift at $\sim N$ (permutations are full-rank). Substituting each near-global head's softmax$(QK^\top)$ with its best per-image rank-10 SVD costs -0.18% on ViT-B/16 ImageNet-21k FT's 29 low-r_{abs} heads and -1.36% on all 68 near-global heads. On DINO, rank-10 at all 117 near-global heads (81% of the model) keeps CLS cos $= 0.9665$ (Table 10). The rank curve at the 68 near-global heads of ViT-B/16 ImageNet-21k FT (full 50k val) is monotone but not broadcast-dominated: k=1 collapses top-1 by -50.7%, k=3 by -14.4%, k=10 by -2.1%, k=50 by -0.03% (Appendix M). Near-global heads therefore behave as a broadcast plus a handful of corrections.

Rank-k also covers mid-rank diffuse-local heads, with backbone-dependent k. The mid-rank DIFFUSE-LOCAL pool (formerly residual heads with eff_rank$_{90} > 40$, redistributed by the rule of Sec. 3.3) sits at different effective ranks across backbones (Table 18). On 4 of the 10 pretrained backbones (DeiT-Small, ViT-B ImageNet-1k, MAE-B, DINO-B) and on MAE-tiny scratch, rank-50 across this pool matches or exceeds the random-head control (CLS cos ≥ 0.9975 or Δ top-1 favorable to the targeted substitution). On MAE-tiny the eff_rank$_{0.9}$ median is 52 and rank-30 already gives CLS cos 0.9993 (Table 17). On ViT-S/16, ViT-B ImageNet-21k FT, DINOv2, CLIP, and SigLIP, rank-50 underperforms the random-head control at that fixed k, but the rank curve is monotone and converges. We verify this on SigLIP, the strongest fail-case (13 such heads; cos 0.945 at k=30 and 0.966 at k=50, against 0.998 and 0.9996 for the random-head control): the substitution reaches cos 0.990 at k=100, 0.999 at k=150, and 1.000 at k=196 (full rank, identity). The other four backbones reach parity at k=30 to 100; see Table 26. The failure at k=50 on these five backbones comes from truncating the rank budget below the backbone-dependent k at which the low-rank structure is captured; these heads remain inside the operator library, at a higher per-backbone k.

F.2 Heads verified lossless in the compression regime

Table 27 lists, per backbone, the groups of heads whose substitution by a fixed-cost operator (shift kernel, per-image broadcast, or rank-10) was tested on its own, and whether the group passed: top-1 within the 95% CI of zero on classifier backbones, CLS cos-sim ≥ 0.95 on encoder-only backbones. Groups are tested separately, and the *Verified* column adds the sizes of the groups that pass. Most

Table 23: Kernel substitution preserves predictions across seven backbones. Top: full ImageNet-1k val, Δ top-1 (%, CI ± 0.4). Bottom: encoder-only backbones, CLS cos to baseline on 9,984 images. Full table with controls in Table 7.

Model	Heads (category)	n_h	Metric
Full ImageNet-1k val (n=50k); Δ top-1 (%):			
ViT-B/16 ImageNet-21k FT	delta-shift	4	$\mathbf{+0.20}$
DeiT-Tiny ImageNet-1k	delta + soft-shift	3	$\mathbf{-0.11}$
ImageNet-1k val subset (n=9,984); CLS cosine sim. to baseline:			
ViT-B/16 MAE	delta + soft-shift	9	$\mathbf{0.9993}$
ViT-B/16 DINO	delta-shift	5	$\mathbf{0.9960}$
ViT-B/14 DINOv2	delta-shift	12	$\mathbf{0.9844}$
CLIP ViT-B/16	delta-shift	6	$\mathbf{0.9915}$
SigLIP ViT-B/16	delta-shift	4	$\mathbf{0.9833}$

Table 24: τ-sweep of the adaptive substitution on ViT-B/14 DINOv2. CLS cos-sim to baseline on 2,048 ImageNet val images, and mean k_h/d_h over the substituted non-shift heads ($k_h \geq 10$). The taxonomy and k_h are computed with the minimal reproducer on an 8-image calibration batch, so the τ=0.9 entry is lower than the 0.896 of Table 2.

τ	CLS cos-sim	mean k_h/d_h
0.90	0.841	0.580
0.93	0.872	0.614
0.95	0.900	0.647
0.97	0.923	0.691
0.99	**0.952**	0.758

cells are the r_{abs}-gated targeted substitutions of Tables 7, 8, 9, and 10. Cells marked † come from an additional run that substitutes jointly every head assigned to the category by the Stage-1 rules on a 32-image ImageNet calibration batch, without the r_{abs} gate, evaluated on 9,984 images (2,048 for DINOv2); their head counts can therefore differ from Table 5. On the four supervised backbones other than ViT-B/16 ImageNet-21k FT, rank-10 applied to all near-global heads at once fails the CI test (-0.77 to -2.28% top-1), and the gated subsets were not tested separately there. Untested heads count as not verified, so the *Verified* column is a lower bound.

G Behavioral mapping by operator class

Re-evaluating full ImageNet-1k val under each substitution at *all* category heads gives per-class profiles that are *distinct* across operators, on both ViT-B/16 ImageNet-21k FT and DeiT-Tiny (Table 11). Substituting shifts leaves per-class accuracy unchanged (no class drops $\geq 10\%$ on ViT-B; per-class std 1.47% = noise floor). Broadcasting all FIXED-TARGET heads damages fine-grained recognition (top-affected: dog breeds, small hand-held tools); rank-10 on all NEAR-GLOBAL heads hurts elongated/articulated objects (stretcher, dumbbell, parallel-bars, accordion). The pairwise Spearman correlation of per-class deltas is small between shifts and the other two ($\rho = 0.09$-0.17) and moderate between broadcast and rank-10 ($\rho \approx 0.32$-0.36): the operators are not interchangeable, and each affects a different set of classes.

H Residual-stream faithfulness

A substitution that preserves top-1 could still be merely output-faithful, with the perturbation absorbed by downstream heads via redundant pathways. We capture the residual stream at every block boundary on a 256-image batch and report mean cos-sim between baseline and substitution per block (Table 12). On ViT-B/16, the validated substitutions stay above 0.97 at every block boundary (shift 0.995, broadcast-low-r 0.986, rank-10-low-r 0.990): the substitution preserves the residual-stream trajectory through the network as well as the final logits. The all-heads versions that damage top-1 also damage the residual stream, with the cos-sim minimum at the layers where the layer-stratified failure localizes. DeiT-Tiny rank-10 is the boundary case: substitution drops both faithfulness (cos-

Table 25: Main broadcast and rank-k substitution results. Broadcast (Proposition 2) at low-r_{abs} FIXED-TARGET heads; rank-10 (Proposition 3) at NEAR-GLOBAL; rank-50 at the mid-rank DIFFUSE-LOCAL pool. Top: Δtop-1 on full val. Bottom: CLS cos to baseline. Full breakdown + controls in Tables 8, 10, 18.

Model	broadcast (low-r FT)	rank-10 (NG, low-r)	rank-50 (mid-rank DL)
Δ *top-1 (%) on full ImageNet-1k val:*			
ViT-B/16 ImageNet-21k FT	-0.15	-0.18	-0.30
ViT-B/16 ImageNet-1k	$+0.08$	-	$\mathbf{+0.01}$
DeiT-Small	$+0.07$	-	$\mathbf{-0.24}$
CLS cosine sim to baseline (n=9,984):			
DINO-B	0.992 (all 14 FT)	0.967 (all 117 NG)	$\mathbf{0.998}$
MAE-B	-	-	$\mathbf{1.000}$
SigLIP-B	-	-	0.945

Table 26: Rank-k recovery curve at the mid-rank diffuse-local heads of the five backbones where rank-50 underperforms the random-head control. Per-image rank-k SVD at all heads of the pool (n). Top: Δtop-1 (%) on 9,984 ImageNet val images (95% CI $\pm$0.80 for ViT-S/16, $\pm$0.69 for ViT-B/16). Bottom: CLS cos to baseline on 9,984 images (512 for DINOv2, whose 1,369 tokens make k=196 less than full rank). *Parity*: smallest k with top-1 in the CI of zero or cos $\geq$ 0.95. On SigLIP the random-head control gives cos 0.998 at k=30 and 0.9996 at k=50; the targeted substitution matches it by k=100.

Backbone	n	k=10	30	50	100	150	196	Parity
ViT-S/16 ImageNet-21k FT	11	-10.66	-2.27	-0.89	-0.04	$+0.11$	$+0.01$	100
ViT-B/16 ImageNet-21k FT	12	-1.54	-0.42	-0.36	-0.24	-0.06	-0.02	30
CLIP ViT-B/16	13	0.943	0.989	0.996	0.999	1.000	1.000	30
ViT-B/14 DINOv2	17	0.606	0.901	0.976	0.995	0.999	0.999	50
SigLIP ViT-B/16	13	0.870	0.945	0.966	0.990	0.999	1.000	50

sim 0.89) and top-1 (-3.65% on clean, Table 16); its near-global heads have higher effective rank than ViT-B/16's, and the mechanism-level probe predicts the top-1 failure.

Out-of-distribution robustness. Under Gaussian noise, blur, and contrast reduction, the substitution-vs.-baseline gap matches the clean gap (ViT-B/16 within $\pm$0.6%, DeiT-Tiny shift/broadcast within $\pm$1.1%, n=9,984; Table 16): the substitution gap does not grow under these corruptions.

I Cross-attention in BLIP captioning

The SOLA diagnostics carry over to cross-attention with one substitution in the indexing: rather than image-patch queries reading image-patch keys (self-attention), the queries are decoder text tokens and the keys are image patches. The image-mean residual r_{abs} then measures how much the per-query attention row varies *across text tokens within a single image*; a low r_{abs} means every text token reads the same set of image patches. The mapping into the SOLA categories is therefore: "FIXED-TARGET" = "every text token attends to the same image patches per image, but those patches change with the image," and the substitution that replaces the head with its row-mean is a *per-image broadcast over text positions*, not over image positions.

On BLIP-base captioning, all 144 decoder cross-attention heads (12 layers $\times$ 12 heads) have median $r_{\text{abs}} \leq 0.34$ and eff_rank$_{90} \in \{1, 2\}$, so they all sit in FIXED-TARGET under this re-reading. Per-image broadcast and per-image rank-10 at *every head* preserve $100/100$ generated captions on COCO val (greedy decoding, max =30 tokens; Table 19). A uniform-attention sanity control breaks all 100 captions (median Jaccard 0.20), confirming that the substitution is actively replacing the cross-attention computation rather than getting bypassed by the decoder's text-side residual stream. Mechanistically, the decoder consumes a single image-dependent visual summary per layer and reads it identically from every text position; the text axis is therefore informational overhead from the cross-attention's perspective. The FIXED-TARGET label in Figure 2 for BLIP reflects this remapped diagnostic, not a self-attention pattern.

Table 27: Heads verified lossless in the compression regime. Each cell gives the number of heads in a tested group: bold if the group passes, in parentheses if it fails, "-" if not tested. Bc.: per-image broadcast. FT: FIXED-TARGET. DL: DIFFUSE-LOCAL. NG: NEAR-GLOBAL. †: category-wide run without the r_{abs} gate (see text).

Backbone	n_{tot}	Shift	Bc. FT	Bc. DL	Rank-10 NG	Verified
DeiT-Tiny ImageNet-1k	36	**3**	**2**	-	(23)†	5 (13.9%)
DeiT-Small ImageNet-1k	72	**4**	**1**	(3)	(42)†	5 (6.9%)
ViT-S/16 ImageNet-21k FT	72	**4**	(3)	(2)	(33)†	4 (5.6%)
ViT-B/16 ImageNet-1k	144	**7**†	**3**	**2**	(86)†	12 (8.3%)
ViT-B/16 ImageNet-21k FT	144	**4**	**5**	**3**	**29**	41 (28.5%)
ViT-B/16 MAE	144	**9**	**11**	**3**†	**98**†	121 (84.0%)
ViT-B/16 DINO	144	**5**	**10**	**2**	**66**	83 (57.6%)
ViT-B/14 DINOv2	144	**12**	(18)	(7)†	**35**	47 (32.6%)
CLIP ViT-B/16	144	**6**	(15)	**3**	**71**†	80 (55.6%)
SigLIP ViT-B/16	144	**4**	(89)	(5)†	(33)†	4 (2.8%)
All ten	1,188					402 (33.8%)

Table 28: COCO results on ViT-B/16 ImageNet-21k FT.

Metric	Value
COCO-Stuff linear probe mIoU, baseline	0.159
COCO-Stuff linear probe mIoU, adaptive substitution	0.161
category flip rate, ImageNet vs. COCO calibration (mean ± std, 8 pairs)	18.5% ± 5.8%
category flip rate, majority vote over 8 batches	12.5%
category flip rate, ±0.05 threshold perturbation (Table 30)	1.3%

J COCO: dense prediction and taxonomy stability

Linear segmentation probe. To test a dense task, we train a linear classifier on the frozen final patch tokens of ViT-B/16 ImageNet-21k FT to predict COCO-Stuff labels at patch resolution (majority label per 16×16 patch; 3,000 train2017 images, 5 epochs, Adam). The same probe is then evaluated on 500 val2017 images twice: through the unmodified backbone, and with the adaptive substitution of Cor. 3.1 (100% coverage) installed at inference, with k_h computed on a 32-image COCO calibration batch. Over the 92 classes present in the evaluation subset, mIoU is 0.159 for the baseline and 0.161 with the substitution (Table 28), so on this probe the substitution does not degrade dense prediction. This is a single run with a low-capacity probe; boundary accuracy at pixel resolution is not measured.

Taxonomy stability across datasets. We recompute the Stage-1 diagnostics and categories of the same backbone on pairs of 32-image calibration batches, one from ImageNet val and one from COCO val2017, for 8 independent pairs. On average 18.5% of the 144 heads change category (std 5.8%, range 8.3-25.7%); with a majority vote over the 8 batches of each dataset, 12.5% change. Of the 213 changes, 112 move a head between FIXED-TARGET and NEAR-GLOBAL, and 97 involve the residual OTHER class before the Stage-2 redistribution. This is an order of magnitude above the flip rate under threshold perturbations (Table 30; mean 1.5%, max 8.3%): the taxonomy depends more on the calibration images than on the thresholds, and we do not claim that it is dataset-invariant.

K Per-head fingerprint diagnostics

K.1 Detection rules and threshold tuning

The structural-category labels of Sec. 3.3 are computed from a small set of per-head diagnostics, evaluated on a 32-image calibration batch using the per-(image, query) attention entries $\alpha_{i,j}^{(h)}$ on the spatial sub-grid (CLS and any register tokens are stripped before recentering).

Quantities.

- **Self-mass**: $\mathbb{E}_{x,i}[\alpha_{i,i}^{(h)}]$, the fraction of attention placed on the query's own position.
- **Escape mass at** K=3: $e_3 = \mathbb{E}_{x,i}\big[1 - \sum_{\|j-i\|_\infty \le 1} \alpha_{i,j}^{(h)}\big]$, i.e. the fraction of attention placed outside the 3×3 window centred on the query. High values mark globally-spread attention.
- **eff_rank$_\tau$ (head-level)**: $\hat{k}_\tau^{(h)}$ from Cor. 3.1; the per-image quantity is $k_\tau^{(h)}(x)$. We use τ=0.9 as the default and write $\mathrm{eff_rank}_{90}$ in tables for brevity (and $\mathrm{eff_rank}_{99}$ for τ=0.99).
- **Per-image argmax-concentration**: maximum, over absolute keys $j^\star$, of the fraction of (image, query) pairs whose attention argmax falls on $j^\star$. High values flag a single key that wins for most queries (the fixed-target signature).
- **Modality**: number of local maxima of the recentered mean kernel $\bar{\kappa}(\delta) = \mathbb{E}_{x,i}[\alpha_{i,i+\delta}^{(h)}]$ above a 0.02 floor.
- **Entropy of** $\bar{\kappa}$ in nats: a perfect delta has entropy 0; a uniform 11×11 kernel has $\log(121) \approx 4.8$.
- **2D Gaussian-fit residual**: ℓ_2 residual of the best 2D Gaussian fit to $\bar{\kappa}$, normalised by $\|\bar{\kappa}\|_2$.
- **TE residual** r: per-(image, query) L_1 deviation between $\alpha_{i,\cdot}^{(h)}$ and the recentered mean $\bar{\kappa}$ (the diagnostic gating Proposition 1, Sec. 3.2).
- **Image-mean residual** r_{abs}: per-image L_1 deviation between $\alpha_{i,\cdot}^{(h)}$ and the per-image row-mean $\bar{A}_X$ (the diagnostic gating Proposition 2, Sec. 3.2).

Detection rules (Stage 1). Each (layer, head) is labelled by applying the rules of Table 29 in order; the first matching rule wins. Heads that match no rule fall into a residual OTHER class that is reassigned in Stage 2 below. The rule order goes from the most structurally restrictive class to the least, so a head that satisfies multiple criteria receives the tighter label. Categories that do not fire on any backbone in our sweep are still listed: the diagnostic admits heads of that type even though the ten ViTs we test never produce one.

Table 29: Structural categories used to label every (layer, head) (Stage 1), listed in the order they are tested (first matching rule applies). Heads matching no rule fall into a residual OTHER class that is reassigned in Stage 2.

Category	Detection rule
fixed-target	per-image argmax-concentration > 0.5
near-global	escape mass at $K=3 > 0.7$
self-only	self-mass > 0.7 AND peak at origin
delta-shift	off-center peak with top-1 mass > 0.5, modality ≤ 2
multi-delta	top-3 mass > 0.7, modality ≥ 2, entropy < 2
soft-shift	off-center peak, $r < 0.4$, modality ≤ 3, entropy < 2.5
gaussian-local	2D Gaussian fit residual < 0.15, modality 1, escape < 0.5
diffuse-local	escape < 0.5

Stage 2: redistribution of OTHER. The OTHER class collects heads that match no Stage-1 rule. Each such head is reassigned by its $\mathrm{eff_rank}_{90}$ to one of the two rank-k categories: OTHER $\to$ NEAR-GLOBAL if $\mathrm{eff_rank}_{90} \le 40$ and $\to$ DIFFUSE-LOCAL otherwise. The threshold 40 is the empirical boundary between the observed maximum $\mathrm{eff_rank}_{90}$ inside Stage-1 NEAR-GLOBAL (39 on ViT-B/16 ImageNet-21k FT) and the observed minimum inside DIFFUSE-LOCAL (40); we set it once on ViT-B/16 ImageNet-21k FT and apply it unchanged to the other backbones. After Stage 2 the inventory has eight categories and no residual class.

Threshold tuning. The Stage-1 thresholds are hand-set rather than grid-searched, with the cascade ordering (most-restrictive-first) doing most of the disambiguation. Cuts were chosen on the ViT-B/16 ImageNet-21k FT calibration batch so that each rule's positive set agrees with the per-(image, query) attention visualisation of the heads in that class, and were then applied unchanged across the remaining nine backbones; per-backbone counts (Appendix Table 5) confirm the rule set transfers. The Stage-2 cut at $\mathrm{eff_rank}_{90}$=40 is the empirical boundary described above. The category labels do not gate substitution: substitution is gated by the proposition-level diagnostic of Sec. 3.2 (r for Proposition 1, r_{abs} for Proposition 2, $\mathrm{eff_rank}_{90}$ for Proposition 3), so threshold sensitivity in Tables 29 only affects the structural inventory, not the substitution-error bounds.

Threshold-robustness ablation. We perturb each Stage-1 threshold by ± 0.05 and re-classify every (layer, head) on each of the ten backbones, reporting the worst-case fraction of heads that change category (Table 30). Across $1{,}188$ heads in ten backbones and eight thresholds, the mean per-backbone flip rate is 1.5% and the maximum is 8.3% (near-global escape-mass cutoff on DeiT-Tiny / DeiT-Small / ViT-S/16, where the 36-72-head total amplifies single-head flips). Three thresholds (self-mass, top-3 multi-modality, Gaussian-fit residual) are entirely insensitive across all backbones. The inventory is therefore stable under modest threshold perturbations, and the per-prop substitution diagnostics (r, r_{abs}, eff_rank$_{90}$) carry no threshold sensitivity at all. It is less stable under a change of calibration images: computing the diagnostics on COCO instead of ImageNet changes the category of 18.5% of the heads of ViT-B/16 ImageNet-21k FT (Appendix J).

Table 30: Threshold-robustness ablation. Worst-case percentage of heads that change category under ± 0.05 perturbation of each Stage-1 threshold (FT$_{\text{arg}}$: fixed-target argmax-conc., NG$_e$: near-global escape mass, m_{self}: self-mass, δ-top1: delta-shift top-1 mass, multi$_3$: multi-delta top-3, soft-r: soft-shift TE residual, gauss: Gaussian-fit residual, DL$_e$: diffuse-local escape mass). Last column is per-backbone mean over the eight thresholds.

Backbone	n	FT$_{\text{arg}}$	NG$_e$	m_{self}	δ-top1	multi$_3$	soft-r	gauss	DL$_e$	mean
DeiT-Tiny	36	2.8%	8.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.4%
DeiT-Small	72	6.9%	8.3%	0.0%	1.4%	0.0%	0.0%	0.0%	0.0%	2.1%
ViT-S/16	72	2.8%	8.3%	0.0%	1.4%	0.0%	0.0%	0.0%	1.4%	1.7%
ViT-B/16 ImageNet-1k	144	6.2%	6.2%	0.0%	0.0%	0.0%	0.0%	0.0%	2.8%	1.9%
ViT-B/16 ImageNet-21k	144	4.2%	3.5%	0.0%	0.0%	0.0%	0.0%	0.0%	2.8%	1.3%
ViT-B/16 MAE	144	4.2%	3.5%	0.0%	1.4%	0.0%	0.7%	0.0%	2.8%	1.6%
ViT-B/16 DINO	144	2.8%	2.8%	0.0%	0.0%	0.0%	0.0%	0.0%	1.4%	0.9%
ViT-B/14 DINOv2	144	4.2%	5.6%	0.0%	0.0%	0.0%	0.0%	0.0%	1.4%	1.4%
CLIP ViT-B/16	144	2.8%	5.6%	0.0%	0.7%	0.0%	0.0%	0.0%	2.1%	1.4%
SigLIP ViT-B/16	144	4.9%	4.9%	0.0%	1.4%	0.0%	0.0%	0.0%	1.4%	1.6%

K.2 Per-head diagnostics on ViT-B/16 shift candidates

All four kernels are concentrated on a single off-center cell (delta-like), not Gaussian blobs; self-mass is at most 0.06 and query-pair correlation is near zero. The detected offsets $\hat{\Delta} = (-1, 0), (0, -1), (1, 0), (0, 1)$ for heads $(1, 10), (2, 6), (2, 3), (3, 1)$ form a complete cardinal basis. Table 31 reports the four diagnostics from Section E.2 and the detected offset $\hat{\Delta}$ for every head we fingerprint in ViT-B and DeiT-Tiny. Self-mass is uniformly low (< 0.11); argmax-concentration is at the noise floor ($\sim 1/n_{\text{queries}} = 1/196$ if heads attended to random absolute keys, here ~ 0.013, indicating heads do *not* fixate on a single absolute key but rather follow the query); query-pair correlation is close to zero (heads are not query-independent); and the recentered fingerprint entropy stays below 2 nats in all cases (a perfectly delta kernel would have entropy 0, a uniform 11×11 kernel would have entropy $\log(121) \approx 4.8$). All seven heads have detected offsets at unit distance from the query, spanning all four cardinal directions.

K.3 Per-image shifts do not appear

Section F found that broadcast heads are per-image (their row changes with the image); a related hypothesis is whether attention is sometimes a sharp delta per image but at a content-dependent offset $\Delta(x)$, a "per-image shift" (translation-equivariant within an image, varying across images). For each head we compute the per-image mean recentered kernel (averaged over interior queries within a single image) and report (i) its top-1 mass, (ii) the std of the per-image peak offset across images, and (iii) the count of distinct peak offsets. A per-image-shift head should have high per-image top-1 *and* multiple distinct peaks. We do not find such a head: on three representative backbones (ViT-B/16 ImageNet-21k FT, DINOv2, MAE), of the heads with per-image top-1 > 0.5 ($n = 7, 14, 10$ respectively), *every one* has a single peak offset across the calibration batch. Heads with high peak-std uniformly have low per-image top-1 (the peaks are argmax-noise on a dispersed distribution). The translation-equivariant heads we observe are always image-independent translations.

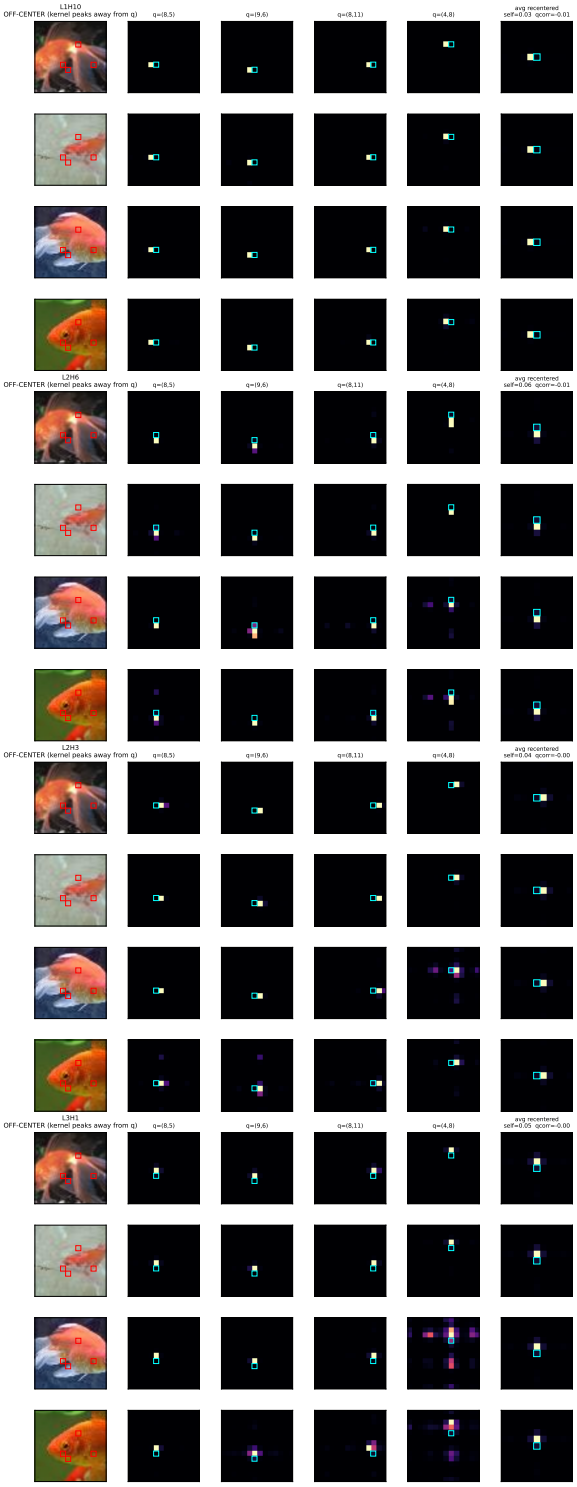

Figure 7: Attention maps and recentered fingerprints for the four ViT-B candidate heads. Rows: heads. Columns: input image (query positions marked), per-query attention maps, and the recentered mean kernel $\bar{\kappa}(\delta)$ averaged over a 32-image batch and all spatial queries.

Table 31: Per-head diagnostics on a 32-image calibration batch (n = 32×196 query/image pairs). Argmax-concentration is the maximum fraction of queries for which the per-(image, query) attention argmax lands on the same absolute key, averaged over the batch. $\hat{\Delta}$ is the recentered-fingerprint argmax.

Model	Head	self mass	argmax conc.	qcorr	entropy	$\hat{\Delta}$
ViT-B/16	L1H10	0.029	0.013	-0.005	0.753	$(-1, 0)$
ViT-B/16	L2H6	0.063	0.013	-0.006	1.385	$(0, -1)$
ViT-B/16	L2H3	0.040	0.012	-0.001	1.470	$(1, 0)$
ViT-B/16	L3H1	0.048	0.017	-0.001	1.849	$(0, 1)$
DeiT-Tiny	L1H2	0.108	0.013	-0.005	1.290	$(0, -1)$
DeiT-Tiny	L2H0	0.094	0.013	-0.006	1.486	$(1, 0)$
DeiT-Tiny	L2H1	0.099	0.013	$+0.001$	1.987	$(0, 1)$

Table 32: SVD spectrum of attention by category (median across heads, 128-image batch). Three regimes: FIXED-TARGET at eff_rank≈ 1 (broadcast); NEAR-GLOBAL at intermediate rank ($\sim$10); DELTA-SHIFT full-rank but a permutation (kernel-substitutable despite high SVD rank).

Model	Category	n_h	$\sigma_1/\|A\|_F$	$\text{rank}_{90\%}$	$\text{rank}_{99\%}$
ViT-B/16 ImageNet-21k FT	fixed-target	52	0.912	1.9	7.7
ViT-B/16 ImageNet-21k FT	near-global	68	0.632	10.4	38.5
ViT-B/16 ImageNet-21k FT	diffuse-local	7	0.236	74.7	127.9
ViT-B/16 ImageNet-21k FT	delta-shift	4	0.114	130.0	161.7
ViT-B/16 DINO	fixed-target	14	0.962	1.4	4.2
ViT-B/16 DINO	near-global	117	0.734	5.6	24.6
ViT-B/16 DINO	delta-shift	5	0.107	157.6	178.6
ViT-B/14 DINOv2	fixed-target	43	0.948	1.6	11.9
ViT-B/14 DINOv2	near-global	67	0.537	17.7	63.2
ViT-B/14 DINOv2	delta-shift	12	0.104	183.4	227.7
CLIP ViT-B/16	fixed-target	59	0.955	1.3	7.7
CLIP ViT-B/16	near-global	63	0.630	15.1	55.9
CLIP ViT-B/16	delta-shift	6	0.113	147.9	176.4
SigLIP ViT-B/16	fixed-target	92	0.963	1.2	5.0
SigLIP ViT-B/16	near-global	27	0.600	17.7	54.5
SigLIP ViT-B/16	delta-shift	4	0.168	132.0	164.5
ViT-B/16 MAE	fixed-target	16	0.969	1.1	3.7
ViT-B/16 MAE	near-global	102	0.586	9.6	40.5
ViT-B/16 MAE	soft-shift	3	0.197	94.6	145.7
ViT-B/16 MAE	delta-shift	6	0.112	143.2	174.2

L SVD spectrum of attention by category

For each (layer, head) we compute softmax$(QK^\top)$ on a 128-image calibration batch and report, per category, the median across heads of (i) $\sigma_1/\|A\|_F$ (top-mode concentration), and the effective rank for (ii) 90% and (iii) 99% of $\|A\|_F^2$. Six backbones in Table 32 (the four supervised backbones not shown follow the same pattern as ViT-B/16 ImageNet-21k FT).

The table also rules out the rank-1 substitution proposal for near-global heads as a salvage of the failed broadcast experiment of Section F: since rank-1 substitution under row-stochastic attention reduces to broadcast, and broadcast fails on near-global, the SVD spectrum (which shows non-trivial energy spread out to $\sim$10 modes) confirms the failure structurally. A rank-10 substitution costs 2.1% top-1 at all 68 near-global heads of ViT-B/16 ImageNet-21k FT (Appendix M). Applying rank-10 factors to V costs about $10Nd_h$ FLOPs per head per block versus $N^2 d_h$ for the unmodified aggregation, a factor $N/10 \approx 20\times$. Realizing this saving requires the factors without forming $A^{(h)}$; the per-image SVD of Proposition 3 forms $A^{(h)}$ first, and we have not implemented a cheaper route.

Table 33: Rank-k recovery curve at all 68 near-global heads of ViT-B/16 ImageNet-21k FT. Per-image rank-k SVD; full 50k ImageNet-1k val.

k	Top-1 (%)	Δ Top-1 (%)
1 (rank-1 = broadcast)	30.34	-50.65
2	55.51	-25.48
3	66.55	-14.44
5	74.16	-6.83
10	78.86	-2.13
20	80.34	-0.65
50	80.96	-0.03

M Rank-k recovery curve at near-global heads

To resolve how much of the rank-10 substitution's success owes to the broadcast component (rank-1) versus the higher SVD modes, we sweep the per-image rank-k SVD substitution across $k \in \{1, 2, 3, 5, 10, 20, 50\}$ at all 68 near-global heads of ViT-B/16 ImageNet-21k FT and measure top-1 on the full 50,000 ImageNet-1k val set (baseline 80.99%, 95% CI $\pm 0.34\%$).

N Wall-clock and energy measurements

This appendix collects the FLOPs accounting, Triton fused-kernel wall-clock measurement, and direct energy measurement referenced in Section 4.3. The measurements cover the shift and broadcast substitutions; the rank-k entry, which needs a per-image SVD, is not accelerated.

FLOPs accounting. A head whose attention is a fixed shift can be implemented as `torch.roll` of the value tensor with appropriate boundary handling. For these heads the per-head Q and K projections, the $QK^\top$ matmul, and the softmax all carry no information and can be skipped, only the V projection is needed for the shifted output. Counting muladd as one operation, an idealized implementation that skips the unnecessary path saves roughly $N(2d\,d_h) + N^2 d_h$ FLOPs per shift head per block, where $d_h = d/n_h$ is the per-head dimension. For ViT-B/16 at 224×224 ($N = 197$, $d = 768$, $h = 12$), each shift head saves about 22 MF, so the four shift heads identified in Section E.2 account for approximately 0.50% of forward FLOPs (1.39% of attention FLOPs). The corresponding figure for DeiT-Tiny's three shift heads is 1.75% of forward FLOPs.

For broadcast-substitutable FIXED-TARGET and DIFFUSE-LOCAL heads, only one row of $QK^\top$ per substituted head per image is needed instead of N rows: the QK matmul drops from $N^2 d_h$ to $N d_h$ FLOPs per head (a factor $N \approx 200$), and the softmax drops by the same factor. Per substituted head per block this is $\approx N^2 d_h \approx 2.5$ MF on ViT-B/16. Cumulative savings:

- ViT-B/16 ImageNet-21k FT: 4 shifts (88 MF) + 5 broadcast fixed-target (12.5 MF) + 7 broadcast diffuse-local (17.5 MF) ≈ 118 MF, $\sim 0.67\%$ of forward FLOPs.
- ViT-B/16 DINO: 5 shifts (110 MF) + 14 broadcast fixed-target (35 MF) ≈ 145 MF, $\sim 0.82\%$ of forward FLOPs.
- DeiT-Tiny: 3 shifts ($\sim 1.75\%$ of forward FLOPs at this smaller $d = 192$); diffuse-local broadcast on its single head adds another $\sim 0.7\%$.

PyTorch reference (no Triton). A naive PyTorch implementation that wraps each substitution as a custom `nn.Module` forward is uniformly slower than baseline despite the FLOPs reductions: the per-head Python orchestration interferes with PyTorch's fused attention and cuBLAS-optimised matmuls. Demonstrating the FLOPs argument as wall-clock therefore requires a fused kernel.

Triton fused SOLA. We implement SOLA in three Triton kernels: (i) shift / kernel substitution (per-head sparse gather of V with on-grid renormalisation), (ii) broadcast (one row of $QK^\top$ per substituted head per image, broadcast to all queries), and (iii) a Triton attention kernel that replaces `F.scaled_dot_product_attention` on the remaining heads. The three are dispatched per head inside a single layer-level forward, so all heads of a layer execute in one launch. The kernels are numerically identical to the PyTorch substitution paths (max difference $< 5 \cdot 10^{-7}$ in fp32 across

Table 34: Triton fused-shift kernel correctness tests. Maximum element-wise difference between the Triton kernel and the pure-PyTorch reference across seven configurations spanning batch size, head count, head dimension, spatial grid, non-spatial token count, kernel mode, and dtype. All differences are at machine epsilon for the corresponding dtype.

B	H_h	D_h	grid	n_{extra}	kernel mode	dtype	$\|\text{Triton} - \text{ref}\|_\infty$
2	6	32	14×14	1	delta	fp32	$0.000 \cdot 10^{0}$
2	6	32	14×14	1	soft	fp32	$2.4 \cdot 10^{-7}$
4	12	64	14×14	1	soft	fp32	$2.4 \cdot 10^{-7}$
4	12	64	37×37	1	soft	fp32	$4.8 \cdot 10^{-7}$
2	12	64	14×14	5	soft	fp32	$2.4 \cdot 10^{-7}$
2	6	32	14×14	1	soft	fp16	$2.0 \cdot 10^{-3}$
2	6	32	14×14	1	random	fp32	$1.8 \cdot 10^{-7}$

the seven test configurations of Appendix O), so all top-1 numbers match the structural substitution exactly. We benchmark the full pipeline (*Triton-FULL* below) end-to-end in fp32 at batch 32, 200 timed runs after 30 warm-ups, on two hardware classes, RTX 3090 (consumer Ampere) and L40S (data-centre Ada Lovelace). Results in Table 13: 1.025-$1.105\times$ end-to-end across the four backbone$\times$GPU settings, with the larger gains on the smaller model (DeiT-Tiny: more shift density relative to total compute) and on the consumer card (3090: weaker Tensor-Core utilisation in baseline). The per-block ratio at the densest substituted layer is variable: positive on ViT-B/3090 (1.10), at parity on ViT-B/L40S, and slightly below on DeiT-Tiny. The e2e win in the Tiny case comes from running every layer through one Triton launch rather than from per-block arithmetic.

Direct energy measurement (RTX 2080 Ti). Wall-clock at parity does not by itself rule out energy savings: fewer FLOPs can lower power draw even when execution time is held constant by latency. As a separate, older study on a single RTX 2080 Ti, we measure energy directly via NVML (`nvmlDeviceGetPowerUsage`) polled at 100 Hz throughout a 60 s sustained inference loop. Idle GPU power is calibrated beforehand and subtracted; we report Joules per image at fixed batch size. Results in Table 14. The pattern is consistent with the FLOPs argument: at fp32 inference on attention-dominated sizes, where softmax$(QK^\top)V$ is a real fraction of forward compute and PyTorch SDPA does not benefit from Tensor Cores, S+B saves 1.3-2.8% of energy per image with a matching 1.8-3.7% reduction in wall-clock time. At fp16, where SDPA dispatches to a heavily Tensor-Core-optimized memory-efficient attention kernel, the substitution loses to the optimized baseline by 2-6%. The fp16 result shows where the FLOPs argument does not translate to wall-clock or energy savings, namely, when the baseline kernel is a hand-tuned memory-efficient fused-attention implementation that already drives the GPU near peak Tensor-Core utilization. On hardware classes without that kernel (older GPUs, mobile NPUs, fp32 research/edge inference) or for kernel-aware deployment frameworks that fuse attention themselves, the structural FLOPs reduction we identify is what an implementation can target.

O Triton fused-shift kernel: correctness tests

The Triton kernel of Appendix N is validated against the pure-PyTorch reference `apply_kernel_to_values` from `kernel_substitution.py` on seven configurations spanning batch size, head count, head dimension, spatial grid, number of non-spatial tokens, kernel mode (delta, peaked-soft, random), and dtype (float32, float16). Maximum element-wise difference:

All differences are at machine epsilon for the corresponding dtype. Top-1 accuracy under Triton substitution exactly matches the structural `kernel_substitution` baseline (85.70% vs 85.70% on DeiT-Tiny at both 224^2 and 384^2, against a baseline of 85.08%).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The structured operator library hypothesis stated in the abstract and Section 1 is supported by the per-substitution and adaptive-coverage experiments in Sections 4.3-4.3 (top-1 preservation, per-class behavioural orthogonality, residual-stream faithfulness, path-patching) on ten ViTs across seven pretrainings, with the SigLIP downstream zero-shot probe in Appendix F.1. Closed-form substitution-error bounds are stated in Sec. 3.2 and proved in Appendix A. The abstract and introduction separate structural coverage, compression, and verified losslessness, and restrict task fidelity at τ=0.9 to the backbones where it holds (Table 2).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations paragraph at the end of Sec. 4 discusses (i) non-universal task-fidelity on high-per-image-variance contrastive backbones (SigLIP, DINOv2), where τ=0.9 is insufficient; (ii) the cross-pretraining causal claim resting on a matched-budget DeiT-Tiny / MAE-Tiny experiment that we do not yet replicate at ViT-B scale; (iii) the modest fp32 wall-clock gain of the Triton implementation, which awaits a fused fp16 kernel for a consistent compute win, and the per-image SVD required by the rank-k entry; (iv) the near-trivial nature of full structural coverage for heads whose k_h approaches N, and the uneven per-backbone fraction of verified lossless heads; (v) the single-backbone tuning of the taxonomy thresholds and the dependence of the taxonomy on the calibration images (Appendix J).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states three substitution-error bounds (Propositions 1-3) and an adaptive layer-level corollary (Cor. 3.1) in Sec. 3.2. Each is proved in Appendix A with explicit assumptions: $A^{(h)}$ row-stochastic and $V^{(h)}$ a fixed value tensor for Props. 1-2 (Hölder + projection-distance argument); the Eckart-Young theorem for Prop. 3; per-image $k_\tau^{(h)}(x) \leq k_h$ for the Cor. 3.1 layer-level bound. Tightness conditions are stated and justified in each proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All models are publicly available timm checkpoints with the exact tags listed (e.g., `vit_base_patch16_224.augreg2_in21k_ft_in1k`, `vit_base_patch14_dinov2.lvd142m`); evaluations are on full ImageNet val. The diagnostics, taxonomies, and substitution rules are defined in Sections 3-3.3 with exact thresholds. A minimal reproducer (`sola_code/`) is included in the supplementary material as a demo, and the full codebase will be released open-source.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same

dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: ImageNet val and the timm pretrained checkpoints are publicly available. The supplementary material includes an anonymised minimal-reproducer code package (`sola_code/`: head diagnostics, the three substitutions, the adaptive Cor. 3.1 forward, ImageNet val loader, and a single-script driver that prints the main numbers on a chosen backbone). The full repository (Triton kernels, behavioral / faithfulness / robustness / energy probes, training scripts) will be released open-source, with each script documented and a top-level `README` mapping commands to paper tables.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Apart from the two matched-budget DeiT-Tiny / MAE-Tiny runs of Sec. 4.2 (100 epochs, AdamW, cosine schedule, 5-epoch warm-up) and the COCO-Stuff linear probe of Appendix J, we evaluate frozen pretrained ViTs from timm on full ImageNet val. Substitution thresholds (TE residual, r_{abs}, SVD spectrum), kernel sizes, calibration batch sizes, and evaluation batch sizes are all specified in Sections 3-F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: 95% binomial confidence intervals are reported on every top-1 substitution result on full ImageNet val (50k images; CI half-width $\sim\pm 0.34$ pp on ViT-B/16). Wall-clock ratios are means over 200 timed forward passes after 30 warm-ups, and energy is averaged over a 60 s sustained inference loop. The taxonomy flip rate across calibration sets is reported as mean $\pm$ standard deviation over 8 batch pairs (Appendix J).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Substitution evaluations ran on a single NVIDIA RTX 3090 (24 GB) or L40S (48 GB) GPU; the Triton fused-kernel wall-clock benchmark covers both classes, reported in Appendix N (Table 13). The earlier energy-measurement study used a single RTX 2080 Ti

(sm_75, 11 GB) and is preserved in Table 14. The from-scratch matched-budget supervised + MAE training runs (Section 4.2) each used a single GPU over 1-2 days. Total compute is on the order of tens of GPU-days.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research analyzes publicly released pretrained models on a publicly available academic dataset (ImageNet), introduces no human-subjects work, releases no high-risk artifacts, and conforms to the Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is mechanistic-interpretability research on existing pretrained vision transformers; it produces no new generative capability or deployment-grade artifact and identifies no specific societal impact distinct from the broader study of vision transformers. The interpretability results may marginally improve diagnosis of trained-model behavior, a positive but indirect effect.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We release no high-risk assets. The released code is analysis machinery for publicly available pretrained ViTs.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use `timm` (Apache 2.0) for pretrained model loading, the timm checkpoints with their original release licenses, ImageNet val (academic-use license), tiny-ImageNet (Stanford academic license), and the PyTorch and Triton libraries (BSD-style and MIT-style respectively). All are cited in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: The new asset is the analysis-and-substitution code repository (taxonomy / diagnostics, kernel / broadcast / rank-k substitutions, Triton fused kernels, behavioral / faithfulness / robustness probes, energy benchmark). The supplementary material contains `sola_code/`, a documented minimal reproducer (4 Python files + README + requirements) covering the head diagnostics and the three substitutions on a single backbone. The full codebase will be released open-source, with per-script docstrings and a top-level `README` mapping commands to paper tables.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves no crowdsourcing and no research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper involves no research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core methodology of this work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.