DIRECT ATTENTION: MULTIPLICATION-FREE QUERY–KEY SCORING WITH RECTIFIED DIFFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention needs a query–key interaction, but that interaction need not be a dot product. We introduce Direct Attention, $s(q,k) = -\sum_r \max(0, q_r - k_r)$, which replaces coordinatewise products with subtraction, rectification, and accumulation while retaining softmax and value aggregation. We derive its exact row-softmax equivalence to coordinatewise minimum, its ℓ_1-distance-plus-key-bias form, and the conditions under which simpler separable scores lose query dependence. A budget-bounded ImageNet-1K conversion of a pretrained ViT-Small updates only Q/K weights and keeps every other parameter fixed. The resulting Direct checkpoint reaches 80.782% top-1 accuracy, compared with 81.392% for the frozen dot-product reference, a difference of -0.610 percentage points under identical official preprocessing. This is one adaptive development trajectory with repeated validation use. In a separate matched Sky130 HD synthesis study, the 16-bit Direct score unit occupies 0.115 times the dot-unit area, giving a Dot/Direct ratio of 8.67. The width sweep links this saving to the different arithmetic primitives; the SRAM-inclusive modeled total area falls by 0.42%. Together, the results identify query–key arithmetic as a practical co-design choice: a rectified difference supports accurate checkpoint conversion and a substantially smaller score unit, with system-level value determined by the surrounding memory and computation.

1 INTRODUCTION

Scaled dot-product attention has become the standard query–key interaction in Transformers (Vaswani et al., 2017). On GPUs, its dense matrix formulation makes effective use of highly optimized multiplication hardware. For an ASIC or FPGA, however, choosing the score also chooses the arithmetic that the circuit must implement. This raises a complementary design question: how much of attention's utility can be retained when the query–key primitive itself changes?

Direct Attention answers this question with a local substitution. It changes the query–key score to

$$s_{\text{DIRECT}}(q,k) = -\sum_{r=1}^{d_k} \text{ReLU}(q_r - k_r), \tag{1}$$

then applies the usual row-wise softmax and weighted sum of values. Figure 1 locates the substitution within the attention pipeline. Each coordinate contributes a one-sided discrepancy: it is zero when the key coordinate meets or exceeds the query coordinate and negative otherwise. The AV product, projections, normalization, and MLP layers retain their existing arithmetic. Throughout this paper, *multiplication-free* refers to query–key scoring.

The central insight is that the score seen by softmax is an equivalence class: adding a query-only constant changes raw scores but leaves attention weights unchanged. Direct's apparently asymmetric discrepancy belongs to the same class as a coordinatewise-minimum score. This observation connects a simple hardware primitive to a familiar similarity construction and gives a precise account of how its geometry differs from dot product.

We connect that analysis to two measurements. In ImageNet-1K checkpoint conversion, Direct reaches 80.782% top-1 against the frozen Dot reference's 81.392%, using Q/K-only updates. In a separate matched circuit comparison, its 16-bit score unit occupies 0.115 times the dot-unit area.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

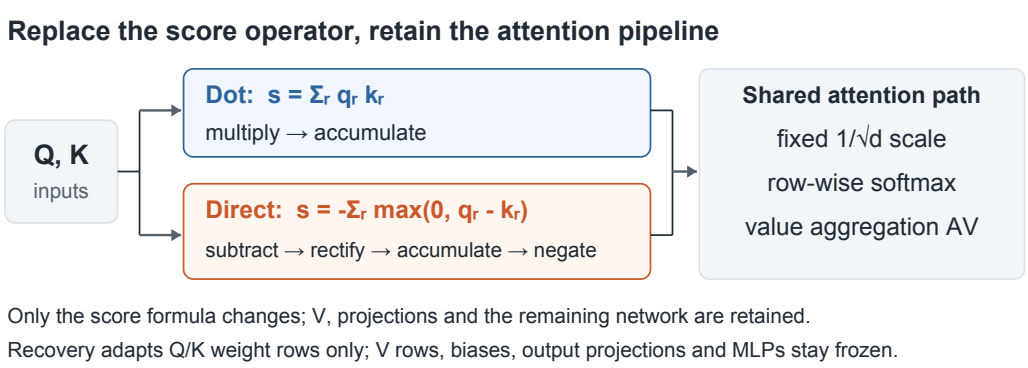


Figure 1: **Change the score primitive; retain the attention pipeline.** Direct replaces each Q/K product with a rectified difference. The fixed score scale, softmax, and value aggregation are shared. During checkpoint conversion, only Q/K weight rows are adapted; the remaining tensors stay fixed. The synthesized score cores cover the highlighted arithmetic, with fixed post-scaling outside the cores.

These are complementary model and circuit measurements, with distinct numerical implementations: CUDA bf16 for accuracy and fixed-point RTL for the score core.

Our contributions are threefold. First, we characterize the rectified- difference score through exact softmax equivalences, invariances, and a separability criterion. Second, we report a complete, budget-bounded Q/K-only ViT conversion and its paired official-preprocessing evaluation. Third, we measure score-unit area across operand widths and place the local savings within a shared logic-and-memory cost model. Transfer and legacy-checkpoint diagnostics in the supplement map where the conversion is effective and where additional adaptation is needed.

2 Related Work

Efficient attention. Linear attention (Katharopoulos et al., 2020), random-feature methods such as Performer (Choromanski et al., 2021), and low-rank methods such as Linformer (Wang et al., 2020) reduce the asymptotic or effective cost of attention. FlashAttention (Dao et al., 2022) preserves exact dot-product attention while reducing memory traffic. These methods target sequence-level computation or data movement; Direct instead changes the scalar Q/K score primitive.

Multiplication-free networks and attention. AdderNet (Chen et al., 2020) and ShiftAddNet (You et al., 2020) replace multiplications in convolutional networks. Closer to our setting, Adder Attention (Shu et al., 2021), multiplication-free piecewise-affine Transformers (Kosson & Jaggi, 2023), ShiftAddViT (You et al., 2023), NoMAD-Attention (Zhang et al., 2024), and EcoTransformer (Gao et al., 2025) demonstrate that attention or Transformer blocks can be built with fewer or no multiplications under particular training and deployment designs. Within this line of work, we focus on the row-softmax equivalence class of the one-sided rectified-difference score, a Q/K-only conversion in which all non-Q/K tensors of a pretrained checkpoint remain frozen, and a matched implementation comparison restricted to the score unit. For nonnegative features, the coordinatewise-minimum representative is the classical histogram-intersection kernel (Swain & Ballard, 1991; Barla et al., 2003; Maji et al., 2008; Vedaldi & Zisserman, 2012); for unconstrained learned Q/K projections we use the more precise term *coordinatewise-minimum score*.

Checkpoint conversion and distillation. Knowledge distillation (Hinton et al., 2015) supplies an optimization signal for adapting the new score geometry. The teacher and student have identical parameter tensors immediately before the student's score operator is replaced; only the student's Q/K projection rows are then permitted to change. This restricted procedure measures how well an inherited representation can accommodate the new score while preserving the rest of the network. It addresses checkpoint conversion; training both operators from a common initialization addresses a different, operator-capacity question.

3 DIRECT ATTENTION

For $Q \in \mathbb{R}^{n_q \times d_k}$, $K \in \mathbb{R}^{n_k \times d_k}$, and $V \in \mathbb{R}^{n_k \times d_v}$, dot-product attention uses

$$S_{ij}^{\mathrm{dot}} = Q_i^\top K_j, \qquad A_{ij} = \mathrm{softmax}_{j \in \mathcal{M}_i}\left(S_{ij}/\sqrt{d_k}\right), \qquad Z_i = \sum_{j \in \mathcal{M}_i} A_{ij} V_j, \tag{2}$$

where $\mathcal{M}_i$ is the set of keys allowed for query i. Direct replaces only the score definition with Eq. (1). It retains the fixed $1/\sqrt{d_k}$ scale. For the tested $d_k = 64$, this is exactly $1/8$; no learned temperature is added during conversion.

3.1 EQUIVALENCE UNDER ROW-WISE SOFTMAX

Proposition 1 (Coordinatewise-minimum equivalence)**.** *For every query Q_i and key K_j,*

$$S_{ij}^{\textsc{Direct}} = \sum_r \min(Q_{ir}, K_{jr}) - \mathbf{1}^\top Q_i. \tag{3}$$

Consequently, Direct is row-softmax equivalent to coordinatewise-minimum scoring. When Q/K coordinates are nonnegative, this representative is the classical histogram-intersection score (Swain & Ballard, 1991; Barla et al., 2003):

$$\mathrm{softmax}_j\left(S_{ij}^{\textsc{Direct}}/\sqrt{d_k}\right) = \mathrm{softmax}_j\left(\sum_r \min(Q_{ir}, K_{jr})/\sqrt{d_k}\right). \tag{4}$$

Proof. The scalar identity $-\max(0, q-k) = \min(q,k) - q$ holds for both $q \le k$ and $q > k$. Summing over coordinates gives Eq. (3); the second term is constant across keys in a query row and cancels exactly in the softmax. □

The equivalence concerns normalized attention, not raw scores. Relative to coordinatewise minimum, Direct's apparent Q/K asymmetry is confined to a query-row constant that softmax removes. Identical forward attention weights do not imply identical training dynamics: at exact $q_r = k_r$ ties, the ReLU and minimum computational graphs may select different valid subgradients (Appendix C).

Proposition 2 (Distance-plus-bias forms)**.** *Modulo query-only row constants,*

$$S_{ij}^{\textsc{Direct}} \overset{\mathrm{softmax}_j}{\equiv} \tfrac{1}{2}\mathbf{1}^\top K_j - \tfrac{1}{2}\|Q_i - K_j\|_1, \tag{5}$$

$$S_{ij}^{\mathrm{dot}} \overset{\mathrm{softmax}_j}{\equiv} \tfrac{1}{2}\|K_j\|_2^2 - \tfrac{1}{2}\|Q_i - K_j\|_2^2. \tag{6}$$

Proof. Apply $-\max(0, q-k) = \frac{1}{2}(k - q - |q-k|)$ coordinatewise. For dot product, expand $\|Q_i - K_j\|_2^2$. The resulting query-sum and query-norm terms are constant over j and disappear under row softmax. □

These decompositions make the distinction precise: Direct combines an ℓ_1 retrieval term with a coordinate-sum key bias, whereas dot product combines a squared-ℓ_2 term with a key-norm bias. This algebraic correspondence does not imply that checkpoints trained with the two scores retrieve the same keys. In particular, pure negative-ℓ_1 attention (Shu et al., 2021; Gao et al., 2025) does not contain Direct's coordinate-sum key bias unless that quantity is constant across keys.

3.2 GEOMETRY, SCALING, AND LEARNING SIGNALS

The changed geometry affects key rankings even in a two-coordinate example. For $q = (1, 1)$ and keys $k_1 = (1, 1)$, $k_2 = (2, 0)$, dot product assigns both keys score 2. Direct assigns 0 and -1: the excess in one coordinate does not compensate for the shortfall in another. Coordinatewise minimum gives 2 and 1, which differ from the Direct scores by the same query-only constant. This example separates the exact Direct–minimum equivalence from the learned relationship between Direct and dot-product attention.

Two additional properties follow directly from the score definition. For a shared coordinate shift $a \in \mathbb{R}^{d_k}$ and a positive scalar c,

$$s_{\text{DIRECT}}(q + a, k + a) = s_{\text{DIRECT}}(q, k), \tag{7}$$

$$s_{\text{DIRECT}}(cq, ck) = c\, s_{\text{DIRECT}}(q, k). \tag{8}$$

The first identity follows because every coordinate difference is unchanged; the second uses the positive homogeneity of ReLU. In contrast, jointly scaling Q and K multiplies dot-product scores by c^2. Thus, with the same fixed softmax scale, Q/K magnitudes affect the concentration of the two attention distributions differently. These are score-level properties; the conversion experiment determines how learned Q/K weights accommodate them within a frozen network.

Away from coordinate ties, the Direct derivatives are $\partial s/\partial q_r = -\mathbf{1}[q_r > k_r]$ and $\partial s/\partial k_r = \mathbf{1}[q_r > k_r]$. Dot-product derivatives instead depend on the opposite operand's value. Direct therefore combines a piecewise-constant local learning signal with cross-key coupling through softmax. At ties, the chosen computational graph determines a valid subgradient. The distinction motivates adapting Q/K after the score swap rather than expecting inherited dot-product coordinates to be immediately optimal.

3.3 Threshold features and separability

The threshold-feature view of histogram intersection (Barla et al., 2003; Vedaldi & Zisserman, 2012) gives, for any bounded signed coordinates $q, k \in [L, U]$,

$$\min(q, k) = L + \int_L^U \mathbf{1}[t \le q]\, \mathbf{1}[t \le k]\, dt. \tag{9}$$

After dropping the key-independent constant $d_k L$, we obtain an inner-product approximation to the coordinatewise-minimum representative score using a finite threshold grid. Small approximation error, however, need not preserve the ranking among top-scoring keys; Appendix G reports both distributional similarity and top-k overlap.

Softmax also exposes a simple failure mode. Suppose a scalar score is additively separable, $\phi(q, k) = f(q) + g(k)$, and every query row has the same allowed-key support $\mathcal{M}$. Then

$$S_{ij} = \sum_r \phi(Q_{ir}, K_{jr}) = F_i + G_j, \qquad A_{ij} = \text{softmax}_{j \in \mathcal{M}}\left(G_j / \sqrt{d_k}\right). \tag{10}$$

Thus every query receives the same distribution over keys. If the mask is query-dependent, rows can still differ because their allowed-key supports differ; Eq. (10) does not remove this mask dependence. Non-separability is necessary to avoid this collapse on shared support, but it does not by itself guarantee a useful learned ranking.

4 ImageNet-1K Q/K-Only Checkpoint Conversion

4.1 Controlled intervention and evaluation

We test whether a pretrained classifier can accommodate Direct by changing its Q/K representation alone. The starting point is the 22M-parameter `vit_small_patch16_224.augreg_in21k_ft_in1k` checkpoint (Steiner et al., 2022), pretrained on ImageNet-21K and fine-tuned on ImageNet-1K with dot-product attention. We copy its tensors into a frozen Dot teacher and a Direct student, then replace all twelve student attention score operators. All additional optimization uses the 1,281,167-image ImageNet-1K training set (Deng et al., 2009).

What changes and what stays fixed. Only the Q and K weight rows may change. V rows, fused-QKV biases, output projections, MLPs, normalization, embeddings, and the classifier remain identical to the source checkpoint. The optimizer receives twelve fused QKV tensors, but gradients on V rows are masked and those rows are restored after each update. Exactly 3,538,944 of 22,050,664 model parameters (16.05%) are semantically trainable. Tensor-identity checks enforce this intervention. This isolates adaptation to Q/K while retaining the pretrained value pathway and downstream parameter tensors.

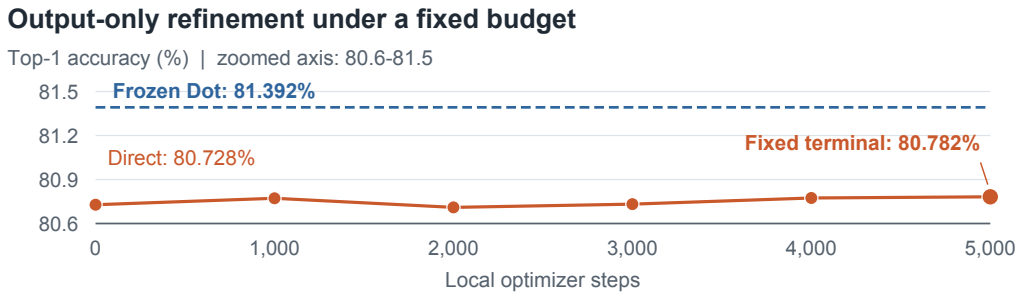


Figure 2: **Output-only refinement within the fixed Phase 3 budget.** The six scheduled validation observations begin from the selected Phase 2 global-step-35,000 checkpoint. The prescribed local-step-5,000 endpoint reaches 80.782%; the horizontal line is the frozen Dot reference. The vertical axis is zoomed. These observations belong to one validation-guided trajectory, not independent training repeats.

Data and numerical controls. Training uses bicubic random-resized 224-pixel crops, horizontal flips, and per-channel mean and standard deviation of 0.5. Both reported checkpoints are evaluated on the same 50,000 validation images with the checkpoint's official configuration: bicubic resize to 248, center crop to 224, and the same normalization. Evaluation uses CUDA bf16 autocast, batch size 128, zero data-loader workers, and the first index of sorted `topk(5)` for top-1. Holding this complete numerical profile fixed reproduces the frozen reference's 40,696 top-1 and 48,073 top-5 correct predictions.

4.2 Budget-bounded conversion protocol

The conversion follows one adaptive trajectory with three recorded phases. For output distributions p_T at temperature T, the output-distillation loss is $\mathcal{L}_{\text{out}} = T^2\text{KL}(p_T^{\text{dot}} \| p_T^{\text{DIRECT}})$. Dense attention matching, $\mathcal{L}_{\text{attn}}$, averages teacher-to-student soft cross-entropy across layers, assigning equal total weight to CLS and non-CLS queries. The combined objective is

$$\mathcal{L} = \alpha\big[(1-m)\mathcal{L}_{\text{out}} + m\mathcal{L}_{\text{attn}}\big] + (1-\alpha)\mathcal{L}_{\text{CE}}, \tag{11}$$

where $\alpha = 0.9$, $T = 3$, and the label term uses smoothing 0.1. Phases 1/2 use $m = 0.5$, AdamW with peak learning rate 10^{-4}, 1,000 warm-up steps, cosine decay, effective batch size 64, and bf16 precision. Phase 1 allows at most one natural training-set pass and uses a patience of three scheduled checks. Phase 2 resumes its terminal state, including optimizer and random-number-generator states, completes any unfinished pass, and permits one additional pass, with a hard global cap of 40,038 updates. After the first pass, five checks without strict improvement stop Phase 2.

Phase 3 starts from the selected Phase 2 checkpoint at global step 35000 and uses output-only distillation ($m = 0$). Its fresh AdamW schedule has peak learning rate 3×10^{-5}, 500 warm-up steps, gradient clipping 1, and zero weight decay. The maximum is 5,000 local updates, or 320,000 training examples. The three phases use deterministic order seeds 42, 43, and 44 as specified in Appendix A. The Phase 2/3 continuation rules were fixed after inspecting Phase 1 and before any Phase 2 parameter update.

Full validation checks occur at the recorded phase boundaries and every 1,000 updates. In Phases 2/3, matching the frozen reference's 40,696 top-1 correct predictions is the target-based early-stop criterion; the phase-specific patience and budget limits above also apply. The trajectory reached the Phase 3 budget, so we report its prescribed local-step-5000 endpoint. Earlier best checkpoints remain provenance records rather than substitutes for that endpoint. Appendix B records the selection rules, evaluation-profile corrections, and restart integrity checks.

Figure 2 shows the final refinement stage. Accuracy changes from 80.728% at its initialization to 80.782% at the prescribed endpoint, with a non-monotonic intermediate trajectory. The narrow range makes the remaining gap to Dot visible: this stage provides a measured refinement, while most of

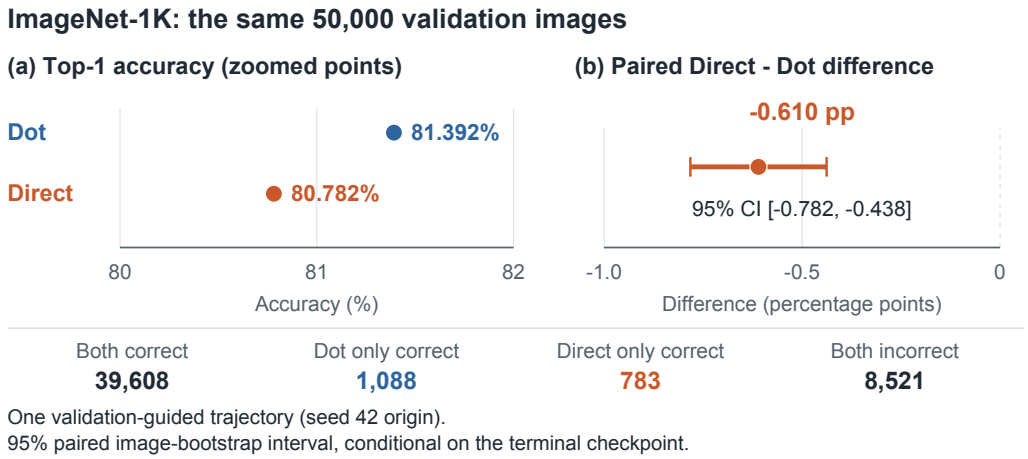


Figure 3: **Official-preprocessing accuracy after Q/K-only conversion.** The points compare the frozen Dot reference with the reported Direct endpoint. The paired 95% bootstrap interval describes image-level variation conditional on this endpoint. Validation was also used during the adaptive training trajectory. Accuracy is measured in CUDA bf16; the hardware figures report a separate fixed-point score-core study.

the conversion has already occurred before its initialization. The experiment thus reports both the achieved endpoint and the budget at which it was obtained.

4.3 Primary result and paired-image analysis

Table 1: **Primary ImageNet-1K result.** Both checkpoints use identical official preprocessing on all 50,000 validation images. Direct is the fixed-budget Phase 3 endpoint of Q/K-only conversion.

Checkpoint	Top-1 (%)	Top-5 (%)	Δ top-1 (pp)
Frozen Dot reference	81.392	96.146	—
Direct (Q/K-only conversion)	80.782	95.870	−0.610

Direct reaches 80.782% top-1 and 95.870% top-5 accuracy, compared with 81.392% and 96.146% for the frozen Dot reference (Table 1). The top-1 difference is -0.610 percentage points. Figure 3 presents the checkpoint comparison and the paired-image uncertainty.

The paired counts make the gap concrete. Both checkpoints correctly classify 39,608 images; Dot alone correctly classifies 1,088, Direct alone 783, and both are incorrect on 8,521. The net difference is therefore 305 images. For each image, we resample its two correctness indicators together, using 10,000 bootstrap resamples with seed 42. The resulting 95% interval for Direct minus Dot is $[-0.782, -0.438]$ percentage points.

This experiment demonstrates that Q/K-only adaptation can support Direct within the pretrained classifier while retaining the entire non-Q/K parameter set. The remaining gap is part of the result. Its statistical scope is a single adaptive development trajectory: repeated validation guided checkpoint retention and continuation, so the paired interval is descriptive, conditional on the reported endpoint. It quantifies neither training-seed variation nor accuracy parity under independent replication. The reference shares the student's source weights and evaluation profile, but remains frozen during adaptation; the comparison therefore measures checkpoint conversion, rather than symmetric operator training.

5 DIAGNOSTIC AND TRANSFER EVIDENCE

The transfer studies test the same intervention in settings with different attention roles. Their value is in exposing which pretrained components accommodate a changed score and which require more extensive adaptation.

Vision-language and translation models. In MiniMind-V (Gong, 2025), applying Q/K-only conversion only to the non-causal vision encoder yields evaluation cross-entropy (CE; lower is better) 1.853 versus 1.841 for the dot baseline, whereas applying it only to the causal decoder yields 2.409. In NLLB-200 (NLLB Team, 2024), among the trained arms sharing a 3,000-step optimizer cap, the encoder-only arm reaches CE 1.9326, versus 1.6976 for the unchanged dot baseline. The decoder-only and cross-attention arms reach 4.7352 and 2.8155, respectively. The smaller vision-encoder gap and larger language-side gaps show sharply different conversion outcomes across the tested components. Module role, sequence statistics, and batch size also vary, so the observed ordering does not isolate causal masking as the mechanism. Within NLLB, equal optimizer-step caps entail unequal effective sample counts; the complete training ledger is reported in Appendix N.

Training and mechanism diagnostics. A short from-scratch DiT study (Peebles & Xie, 2023) reports final training losses 0.1926 for Direct and 0.1971 for dot product. Its short training horizon and lack of Fréchet Inception Distance (FID) evaluation limit interpretation to the training objective rather than generation quality. Full protocols appear in Appendices M, N, and S. Separately, a legacy ViT checkpoint supplies the retrieval and explanation diagnostics in the supplement: 11 of 12 layers pass the alignment criterion, while finite-threshold ranking and the tested interpolation path fail their criteria. Those measurements describe that legacy pair, not the primary checkpoint in Table 1. Together, the studies motivate operator-aware adaptation and distinguish successful classification from identity of the underlying attention mechanisms.

6 SCORE-UNIT HARDWARE STUDY

6.1 MATCHED SCORE CORES AND ARITHMETIC CONTRACT

We implement dot-product and Direct score units under the same synthesis flow. Each unit has 16 parallel lanes and processes a 64-coordinate score in four cycles; six units serve the six attention heads. At operand width W, both units accept signed two's-complement inputs and use the same signed $2W$-bit accumulator, with wrapping rather than saturation. Dot forms a $2W$-bit product before accumulation. Direct computes the difference in $W + 2$ bits, rectifies it, and accumulates its negative. The primary comparison uses $W = 16$ and a 32-bit accumulator. Fixed post-scaling, softmax, AV, projections, and MLPs sit outside these cores.

Bit-exact simulation covers both operators at 4, 8, 16, and 32 bits. For each operator–width pair, 2,007 edge/random vectors and 57,600 shared ImageNet Q/K trace vectors match the integer reference. The traces come from one frozen-Dot checkpoint under the legacy generic preprocessing and are fed to both units. The six-head score tile completes $197 \times 197 \times 4 = 155236$ cycles in RTL simulation. Thus the circuit comparison holds the interface, vector population, accumulator contract, and processing schedule fixed.

Sky130 HD synthesis (SkyWater PDK Authors, 2020) supplies mapped cell areas. The scripts identify the TT library at 25 °C and 1.8 V. A separate sum over the retained 16-bit netlists' cell footprints reproduces the area ratio, as detailed in Appendix O. The measurements characterize this cell library and flow; the exact historical producing-run tool and library identities are incompletely retained.

6.2 WIDTH SCALING AND SYSTEM-AREA CONTEXT

Direct replaces the partial-product array with subtraction and rectification. In the illustrative static-CMOS model, this changes per-coordinate gate-count growth from quadratic to linear in operand bit width. Figure 4 shows the corresponding mapped-area sweep. Accumulation and control are included in the mapped score units, so their areas need not follow a pure asymptotic law. The distinction concerns bit width: dense scoring still requires $\Theta(n^2 d_k)$ coordinate operations for sequence length n.

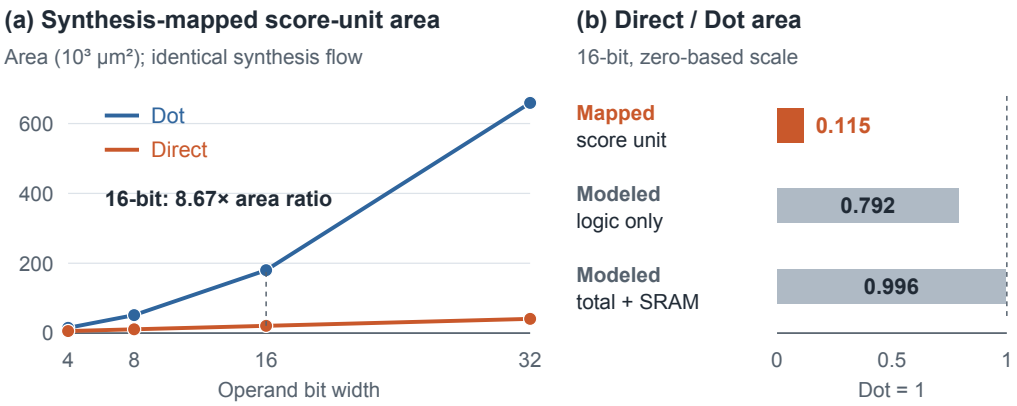


Figure 4: **The arithmetic advantage grows with operand width.** Left: recorded mapped score-unit area at 4, 8, 16, and 32 bits under the shared Sky130 HD flow. Right: Direct/Dot area ratios at 16 bits, separating score-unit synthesis from the modeled logic and SRAM-inclusive total. The local 8.67 Dot/Direct area ratio becomes a 0.42% modeled total-area reduction when unchanged memory and compute blocks are included.

Table 2: Matched 16-bit score-unit mapped area and analytical system-area context. The first row is a score-unit synthesis result; the remaining rows are modeled ledgers with identical non-score blocks and SRAM assumptions. The comparison measures area; timing and energy are separate quantities.

Quantity	Dot	Direct	Direct/Dot
Score-unit area (μm^2)	180135.3	20784.9	0.115
Modeled logic area (mm^2)	4.601	3.645	0.792
Modeled total area incl. SRAM (mm^2)	228.772	227.816	0.996

At 16 bits, the Direct score unit uses 20784.9 μm^2 versus 180135.3 for Dot, or 0.115 times its area (Table 2). The Dot/Direct ratio increases from 2.58 at 4 bits to 4.75 at 8 bits, 8.67 at 16 bits, and 16.2 at 32 bits. This width dependence is the principal circuit-level benefit of changing the scoring primitive.

The system ledger answers a second question: how much of that local saving survives when the rest of the accelerator is held fixed? It assumes identical on-chip weight and activation SRAM, projections, value aggregation, and MLP resources. Modeled logic area decreases from 4.601 to 3.645 mm^2, while the SRAM-inclusive total changes from 228.772 to 227.816 mm^2, a 0.42% reduction. The ledger therefore makes the dominant constraint visible: arithmetic specialization affects one component, whereas unchanged memory occupies most of the modeled system.

Both analytical schedules assign 36508704 cycles per image; the 155236-cycle score tile is the cycle-exact RTL measurement. Historical timing and power records are excluded because their flow lacks demonstrated clock-tree implementation and uses a defective activity conversion. Appendix O documents these records and the retained area evidence. End-to-end timing and energy require a complete routed design and valid activity annotation.

7 Discussion and Conclusion

What the results establish. Direct provides a concrete alternative to coordinatewise products in normalized Q/K scoring. Its algebra explains the resulting similarity, and Q/K-only adaptation yields 80.782% ImageNet-1K top-1 with the rest of the pretrained classifier fixed. The matched score-core study independently establishes a large, width-dependent mapped-area reduction. The common

design lesson is to choose the scoring primitive together with its numerical representation and target hardware.

What determines the trade-off. The remaining top-1 difference of -0.610 percentage points is a measured cost of the reported conversion. Its acceptability depends on deployment requirements; this study specifies no non-inferiority margin. Likewise, the 8.67 Dot/Direct local area ratio and the 0.42% modeled total-area reduction answer different questions. A useful accelerator-level trade-off depends on memory placement, the share of score computation, and the unchanged multiply-intensive layers. The bf16 model evaluation and fixed-point circuit study are separate evidence streams rather than a measured end-to-end accuracy–efficiency frontier.

Scope and next tests. The ImageNet result covers one adaptive trajectory with reused validation; independent seeds and a training-only selection split would assess repeatability. A matched Dot Q/K self-distillation control would isolate the effect of adaptation, while common-initialization Dot/Direct training would test operator capacity. Calibrated negative-distance and learned-temperature baselines would further resolve the roles of score geometry and scale. The retained legacy and transfer diagnostics are checkpoint- and task-specific. On hardware, full routed-accelerator timing and energy remain the next measurements needed to connect score-core savings to deployment.

Direct Attention shows that a small change in the local scoring rule can preserve substantial pretrained classification capability and replace an expensive circuit primitive. Its contribution is the connection between that rule, the adapted representation, and the measured score-unit cost: *change the query–key arithmetic, and quantify where the benefit appears.*

AI USE STATEMENT

Generative AI tools were used for manuscript restructuring and copyediting in English and Chinese; English–Chinese translation; literature discovery; checking mathematical claims and proofs; methodology and experimental design; implementation of code for experiments, analysis, and verification, including the preprocessing audit and corrective training/evaluation pipeline; interpretation of results; consistency and claim–evidence auditing; and LaTeX and table formatting. They were not used to fabricate data or measurements. The authors verified candidate citations against primary publication sources, inspected and executed AI-assisted code, checked reported numbers against saved logs and machine-readable artifacts, and reviewed the final text. The authors take responsibility for all content and artifacts produced with AI assistance.

ETHICS STATEMENT

This work collects no new human-subject data. It uses established public benchmarks and therefore inherits their documentation, representation, privacy, and licensing limitations. The proposed arithmetic could support more resource-efficient inference, but a score-unit estimate alone does not establish lower lifecycle energy or environmental impact. No deployment or safety-critical use is evaluated.

REPRODUCIBILITY STATEMENT

The appendix records data transforms, optimization objectives, seeds, checkpoint-selection rules, integrity checks, statistical populations, and the hardware-flow scope. Numeric manuscript values are drawn from versioned machine-readable evidence and LaTeX macros. The ImageNet conversion and evaluation runs produced and retained configurations, checkpoint hashes, logs, and per-example predictions; these artifacts are not embedded in the PDF. Other archival diagnostics do not uniformly retain the same metadata, and their claims are scoped accordingly. The hardware appendix reports the tested bit widths, named cell library and tools, timing assumptions, and activity source. Reproducing the ImageNet experiment also requires authorized access to ImageNet and a compatible PyTorch/CUDA environment; the appendix specifies the methodological settings but not a fully pinned machine image.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020. doi: 10.18653/v1/2020.acl-main.385.

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

Tutu Ajayi, Vidya A. Chhabria, Mateus Fogaça, Soheil Hashemi, Abdelrahman Hosny, Andrew B. Kahng, Minsoo Kim, Jeongsup Lee, Uday Mallappa, Marina Neseem, Geraldo Pradipta, Sherief Reda, Mehdi Saligane, Sachin S. Sapatnekar, Carl Sechen, Mohamed Shalan, William Swartz, Lutong Wang, Zhehong Wang, Mingyu Woo, and Bangqi Xu. Toward an open-source digital flow: First learnings from the OpenROAD project. In *Proceedings of the 56th Annual Design Automation Conference 2019*, DAC '19, pp. 1–4. Association for Computing Machinery, June 2019. doi: 10.1145/3316781.3326334. URL https://doi.org/10.1145/3316781.3326334. Invited paper.

Annalisa Barla, Francesca Odone, and Alessandro Verri. Histogram intersection kernel for image classification. In *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, volume 3, pp. III–513–III–516. IEEE, 2003. doi: 10.1109/ICIP.2003.1247294. URL https://doi.org/10.1109/ICIP.2003.1247294.

Walid Bousselham, Angie Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. LeGrad: An explainability method for vision transformers via feature formation sensitivity. In *IEEE/CVF International Conference on Computer Vision*, pp. 20336–20345, 2025.

Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 839–847, 2018.

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

Hanting Chen, Yunhe Wang, Chunjing Xu, Boxin Shi, Chao Xu, Qi Tian, and Chang Xu. AdderNet: Do we really need multiplications in deep learning? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations (ICLR)*, 2021.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Rachel Lea Draelos and Lawrence Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020. URL https://arxiv.org/abs/2011.08891. Revised 2021.

Alexandre Englebert, Sédrick Stassin, Géraldin Nanfack, Sidi Ahmed Mahmoudi, Xavier Siebert, Olivier Cornu, and Christophe De Vleeschouwer. Explaining through transformer input sampling. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 806–815, 2023. doi: 10.1109/ICCVW60793.2023.00088. URL https://doi.org/10.1109/ICCVW60793.2023.00088.

Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. In *British Machine Vision Conference*, 2020.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7457–7476, June 2023. doi: 10.1109/TPAMI.2022.3218275. URL https://doi.org/10.1109/TPAMI.2022.3218275.

Xin Gao, Xingming Xu, Shirin Amiraslani, and Hong Xu. EcoTransformer: Attention without multiplication, 2025. URL https://arxiv.org/abs/2507.20096. arXiv preprint, version 2.

Jingyao Gong. MiniMind: The mini and efficient language models for LLM research. https://github.com/jingyaogong/minimind, 2025. MiniMind-V: SigLIP2 vision encoder with a decoder-only LLM backbone.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL https://arxiv.org/abs/1503.02531.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, 2020.

Atli Kosson and Martin Jaggi. Multiplication-free transformer training via piecewise affine operations. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8208–8223. Curran Associates, Inc., 2023. doi: 10.52202/075280-0360. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/19df21cd4931bd0caaa4d8480e9a59cd-Paper-Conference.pdf.

Subhransu Maji, Alexander C. Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2008. doi: 10.1109/CVPR.2008.4587630. URL https://doi.org/10.1109/CVPR.2008.4587630.

NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846, 2024. doi: 10.1038/s41586-024-07335-x. URL https://doi.org/10.1038/s41586-024-07335-x.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Han Shu, Jiahao Wang, Hanting Chen, Lin Li, Yujiu Yang, and Yunhe Wang. Adder attention for vision transformer. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19899–19909. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a57e8915461b83adefb011530b711704-Paper.pdf.

SkyWater PDK Authors. SkyWater SKY130 PDK Documentation. https://skywater-pdk.readthedocs.io/en/main/, 2020. Accessed 2026-09-07.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=4nPswr1KcP.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.

Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991. doi: 10.1007/BF00130487. URL https://doi.org/10.1007/BF00130487.

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

Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012. doi: 10.1109/TPAMI.2011.153. URL https://doi.org/10.1109/TPAMI.2011.153.

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

C. Wolf and Johann Glaser. Yosys – a free Verilog synthesis suite. In Gerald Hilber, Timm Ostermann, and Andreas Rauchenecker (eds.), *Austrochip Workshop on Microelectronics 2013*, pp. 47–52, October 2013. URL https://yosyshq.net/yosys/files/yosys-austrochip2013.pdf.

Junyi Wu, Weitai Kang, Hao Tang, Yuan Hong, and Yan Yan. On the faithfulness of vision transformer explanations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10936–10945, 2024.

Haoran You, Xiaohan Chen, Yongan Zhang, Chaojian Li, Sicheng Li, Zihao Liu, Zhangyang Wang, and Yingyan Lin. ShiftAddNet: A hardware-inspired deep network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Haoran You, Huihong Shi, Yipin Guo, and Yingyan Lin. ShiftAddViT: Mixture of multiplication primitives towards efficient vision transformer. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 33319–33337. Curran Associates, Inc., 2023. doi: 10.52202/075280-1447. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/69c49f75ca31620f1f0d38093d9f3d9b-Paper-Conference.pdf.

Tianyi Zhang, Jonah Yi, Bowen Yao, Zhaozhuo Xu, and Anshumali Shrivastava. NoMAD-Attention: Efficient LLM inference on CPUs through multiply-add-free attention. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 112706–112730. Curran Associates, Inc., 2024. doi: 10.52202/079017-3581. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ccda3c632cc8590ee60ca5ba226a4c30-Paper-Conference.pdf.

A IMAGENET Q/K CONVERSION: TRAINING AND VALIDATION PROTOCOL

Primary conversion study. The primary ImageNet experiment asks how much of a pretrained Dot model's accuracy can be retained when every query–key score is replaced by Direct and adaptation is restricted to Q/K weights. Table 1 reports the main comparison; Table A.1 below gives the expanded validation results. The study followed one budget-bounded, sequential knowledge-distillation (DK) trajectory initialized with seed 42, rather than a seed sweep. It combined dense attention matching in Phases 1/2 with output-only distillation in Phase 3; the phase-specific deterministic order seeds and stopping rules are stated below. The reported endpoint is the Phase 3 terminal checkpoint at local optimizer step 5000, initialized from the selected Phase 2 source at global optimizer step 35000.

Data, initialization, and trainable parameters. Training uses all 1,281,167 ImageNet-1K training images with bicubic random-resized 224-pixel crops, horizontal flips, and per-channel mean and standard deviation 0.5; evaluation uses bicubic resize to 248, center crop to 224, and the same normalization. Immediately before score replacement, the teacher and student copies have identical parameter tensors from the `vit_small_patch16_224.augreg_in21k_ft_in1k` checkpoint used in Table 1. Those tensors were produced with Dot throughout the source ImageNet-21K pretraining and full-model ImageNet-1K fine-tuning; Direct did not participate in either stage. The Dot teacher receives no new updates. In the student, every attention score is replaced by Direct and only the Q/K rows of the fused QKV weights are optimized on ImageNet-1K; V rows, biases, output projections, MLPs, normalizers, embeddings, and the classifier remain frozen and are checked for identity. The experiment tests restricted conversion of a Dot-trained checkpoint, rather than operator capacity under compute-matched, symmetric training. The implementation places 12 fused QKV tensors (5,308,416 scalars) in the optimizer, masks every V-row gradient, and restores all V rows after each update; only 3,538,944 of 22,050,664 model parameters (16.05%) are permitted to change.

Objectives and phase-specific schedules. Let p^{dot} and $p^{\textsc{Direct}}$ denote the two output distributions at temperature T, and let $a^{\mathrm{dot}}_{\ell bhqk}$ and $a^{\textsc{Direct}}_{\ell bhqk}$ denote their attention probabilities. With no additional temperature applied to the normalized attention, and with c the CLS query, the optimized objective is

$$\mathcal{L}_{\mathrm{out}} = T^2\,\mathrm{KL}(p_T^{\mathrm{dot}} \| p_T^{\textsc{Direct}}), \tag{A.1}$$

$$\mathcal{L}_{\mathrm{attn}} = -\frac{1}{L}\sum_{\ell=1}^{L}\left[\beta\,\mathbb{E}_{b,h}\sum_k a^{\mathrm{dot}}_{\ell bhck}\log a^{\textsc{Direct}}_{\ell bhck} + (1-\beta)\,\mathbb{E}_{b,h,q\neq c}\sum_k a^{\mathrm{dot}}_{\ell bhqk}\log a^{\textsc{Direct}}_{\ell bhqk}\right], \tag{A.2}$$

$$\mathcal{L} = \alpha\left[(1-m)\mathcal{L}_{\mathrm{out}} + m\mathcal{L}_{\mathrm{attn}}\right] + (1-\alpha)\mathcal{L}_{\mathrm{CE}}. \tag{A.3}$$

In Phases 1/2, we set $T = 3$, $\alpha = 0.9$, $m = 0.5$, $\beta = 0.5$, and label smoothing to 0.1. AdamW uses learning rate 10^{-4}, 1,000 warm-up steps, cosine decay, gradient clipping 1.0, effective batch 64 (micro-batch 16×4), and bf16 mixed precision. The following records the adaptively specified, budget-bounded protocol. Phase 1 permits at most one natural ImageNet-1K pass (20,019 optimizer steps), evaluates every 1,000 steps, and stops after three checks without improvement. After inspecting Phase 1's intermediate official-validation trajectory, but before any Phase 2 update, we fixed the following continuation and fallback rules. From the Phase 2 boundary onward, all validation uses the archival CUDA bf16 profile: batch size 128, zero data-loader workers, and top-1 taken from the first class index returned by sorted `topk(5)`. This applies to the frozen reference, checkpoint selection, stopping, Phase 3, and final evaluation. Appendix B records the zero-update preflight corrections and the terminal-reporting amendment. If the Phase 1 terminal checkpoint remains below 40,696 correct predictions out of 50,000 (81.392%), Phase 2 resumes that terminal state—not the best retained state—including its AdamW moments and random-number-generator states. It first completes any unfinished suffix of the seed-42 sample order without restarting the learning-rate schedule, then permits at most one new complete ImageNet-1K pass. The new pass uses deterministic order seed 43 and restarts the same 1,000-step warm-up/cosine schedule; the hard global limit is 40,038 optimizer steps. Full validation is performed every 1,000 global steps and at each pass end. Only checks after step 20,019 consume a patience counter: five consecutive checks without a strict improvement in the exact number correct stop Phase 2, and any strict improvement resets the counter. The strongest sub-target checkpoint is retained independently of the rolling terminal state. Phase 2 stops immediately at the first check with at least 40,696 correct predictions.

If bounded Phase 2 does not reach the threshold, one fallback declared at the Phase 1/2 boundary compares common archival-profile full-validation counts for the Phase 1 retained checkpoint, replayed at Phase 2 initialization, and the independently retained Phase 2 best; ties select Phase 1. Before any update, we re-evaluate the selected source once more on the exact 50,000-image validation set. Phase 3 uses a fresh AdamW optimizer, output-only distillation ($m = 0$, $T = 3$, $\alpha = 0.9$, hence $\mathcal{L} = 0.9T^2\,\mathrm{KL}(p_T^{\mathrm{dot}} \| p_T^{\mathrm{DIRECT}}) + 0.1\mathcal{L}_{\mathrm{CE}}$), label smoothing 0.1, peak learning rate 3×10^{-5} with 500 warm-up steps and cosine decay, effective batch 64, bf16 mixed precision, gradient clipping 1, and zero weight decay. It consumes at most the first 320,000 examples of a deterministic seed-44 order (data-loader generator seed 1,000,044; 5,000 optimizer steps), with full validation before training and at local steps 1,000 through 5,000. It likewise stops at the first exact count of at least 40,696. If it remains below that count at the declared bound, the exact local-step-5,000 endpoint is retained, authenticated, and reported without extending the budget or substituting an earlier best checkpoint. The completed run reached this fixed budget without meeting the reference; its terminal endpoint, not an earlier best, is reported below.

Table A.1: Expanded primary ImageNet validation results, corresponding to Table 1. Both rows use the checkpoint's official preprocessing on the same 50,000 images; Direct is the declared fixed-budget Phase 3 endpoint.

Checkpoint	Top-1 (%)	Top-5 (%)	Direct − Dot (pp)
Frozen dot-product reference	81.392	96.146	—
Phase 3 output-only Q/K DK	80.782	95.870	-0.610

Primary outcome and interpretation. Direct obtains 80.782% top-1 and 95.870% top-5 accuracy, with a top-1 difference of -0.610 percentage points from the frozen Dot reference. The conditional 95% paired-bootstrap interval for Direct-minus-Dot top-1 is $[-0.782, -0.438]$ percentage points. Because the same validation set is repeatedly used for checkpoint retention, continue-or-stop decisions, and reporting, this result describes the reported terminal checkpoint from an adaptively continued trajectory, conditional on that checkpoint rather than variation across training seeds. The authenticated fixed-budget Phase 3 endpoint remains below the frozen Dot point reference under the checkpoint-resolved official evaluation. Reaching or missing this point on the repeatedly reused validation split is a single development-trajectory outcome, not evidence of accuracy parity or non-inferiority. The 81.392% target is the frozen Dot reference in this paper, not an external state-of-the-art benchmark.

Relation to the legacy diagnostics. The retrieval, intervention, and explainability studies analyze the frozen bias-preserving legacy checkpoint pair documented in Appendix D, not the primary conversion endpoint. Here, *Stage 3* denotes that historical inference audit and is distinct from training *Phase 3* above. The hardware study uses shared frozen-Dot traces, as specified in Appendix O; it is not a hardware evaluation of either Direct checkpoint.

B Evaluation Populations and Statistical Design

Evaluation populations. The primary conversion comparison uses the complete 50,000-image official ImageNet-1K validation set, with identical preprocessing and numerical settings for the frozen Dot reference and converted Direct endpoint. The legacy Stage 3 diagnostic audit uses separate predefined sample lists: the complete 50000-image ImageNet-1K validation set; 5000 fixed geometry images (five per class); a disjoint 1000-image intervention set (one per class); and all 4580 valid ImageNet-S masks (Gao et al., 2023). Nine empty masks are excluded explicitly. The complete and jointly-correct localization populations are both reported.

Adaptive trajectory and statistical interpretation. The primary ImageNet result is the terminal checkpoint of one Q/K conversion trajectory initialized with seed 42, with the phase-specific order seeds in Appendix A. Scheduled official-validation checks govern checkpoint retention and continuation or stopping. The paired comparison therefore describes this adaptively developed endpoint; repeated validation use and the absence of independent seed repeats preclude a confirmatory or seed-level interpretation. Such a comparison would require a training-only selection split, a prespecified seed-level estimand, and once-only final validation. A symmetric operator-capacity study would also require trained Dot and Direct arms with matched initialization, data, and optimization. The separate legacy endpoint audit and its incomplete training provenance are documented in Appendix D.

Protocol formation and run integrity. At the Phase 1/2 boundary, an exact frozen-Dot gate found a protocol mismatch. Both evaluations used CUDA bf16 autocast, but the attempted launch used evaluation batch size 64 and `argmax`, whereas the archival profile uses batch size 128, zero data-loader workers, and takes top-1 from the first class index returned by sorted `topk(5)`. The launch failed at this gate and performed zero optimizer steps. On the same machine, replaying the archival profile reproduced exactly 40,696/50,000 frozen-Dot top-1 and 48,073/50,000 top-5 correct predictions; the SHA-256 hashes of the resulting per-example top-1 and top-5 correctness vectors matched the stored hashes. A subsequent execution ended during the pre-training bytewise integrity scan while its initialization record still read `initializing`. It produced no run manifest, sample-order artifact, checkpoint, metric or validation record, optimizer update, or model outcome. We archived this zero-update initialization attempt and relaunched the unchanged, hash-pinned protocol from the same frozen Phase 1 terminal state; no partial scan state was reused.

Before any recorded Phase 2 optimizer update and without observing any recorded Phase 2 model outcome, we amended only the terminal-reporting rule. The declared model initialization, trainable tensors, objective, sample orders, optimizer schedules, validation cadence, stopping rules, and compute bounds were unchanged. The amendment required authentication and reporting of either terminal outcome from the already-declared Phase 2/3 sequence. The 81.392% reference remained an early-stop target, not a publication filter: a budget-exhausted Phase 3 run had to report its exact local-step-5,000 endpoint without substituting an adaptively retained best miss. The completed run followed this fixed-budget reporting path.

Checkpoint-load recovery. After the completed validation and checkpoint writes at global step 23,000, Phase 2 stopped because restricted checkpoint loading rejected a `TorchVersion` object stored as software-version metadata. We archived the interrupted run and repaired that metadata by storing version values as plain strings, with corresponding updates to the code-provenance records. CPU-only checks verified that model weights, AdamW state, random-number-generator states, sample order and cursor, validation history, objective, and stopping state were unchanged. A recovery job was launched using that exact step-23,000 checkpoint, retaining the original global limit of 40,038 updates and patience of five checks. This is a continuation of the same development trajectory, not a new seed or a fresh training budget. Intermediate validations from the interrupted run do not replace the authenticated terminal result.

Statistical procedure. For the legacy fixed-checkpoint Stage 3 metric families, image-level intervals use 10000 paired resamples, with Benjamini–Hochberg correction applied separately within each predefined metric family. The primary conversion interval uses the same image resampling, conditional on its reported terminal checkpoint; it is not a seed-level interval. Table B.1 summarizes the populations and their statistical roles.

Table B.1: Evaluation populations and statistical design for Stage 3. The table lists the complete ImageNet-1K validation population, fixed and disjoint geometry/intervention samples, all valid ImageNet-S masks, and the paired-bootstrap and Benjamini–Hochberg units used for each claim family.

Population / analysis	N	Selection	Statistical role
ImageNet-1K validation	50000	complete validation set	accuracy and logit evaluation
Retrieval geometry	5000	five fixed images per class	layer/head geometry
Interventions	1000	one disjoint image per class at offset five	path, components, patch removal
ImageNet-S	4580	all valid foreground masks	localization and controls
Paired bootstrap	10000	paired resamples	95% confidence intervals
Multiple testing	per metric family	Benjamini–Hochberg	reported BH q-values
Geometry/intervention overlap	0	predefined disjoint manifests	threshold calibration isolation
Empty ImageNet-S masks	9	recorded exclusions	not silently discarded

C Direct–Intersection Equivalence and Backend Audit

Equivalent score representatives. The implementation audit compares four representatives:

$$z_1(q,k) = -\sum_r \mathrm{ReLU}(q_r - k_r), \qquad z_2(q,k) = \sum_r \min(q_r, k_r) - \sum_r q_r, \tag{C.1}$$

$$z_3(q,k) = \sum_r \min(q_r, k_r), \qquad z_4(q,k) = \tfrac{1}{2}\sum_r k_r - \tfrac{1}{2}\|q-k\|_1. \tag{C.2}$$

The first two are identical raw scores; the others differ only by a query-dependent row constant. Consequently, the hard gate concerns row-centered scores, attention, attention output, logits, and predictions.

Tie-sensitive gradients. At exact $q_r = k_r$ ties, the audited rectifier and minimum operations choose different valid subgradients. Raw-gradient equality therefore fails at ties while the off-tie and tie-broken checks pass. This does not affect the forward-equivalence claim, but it prohibits a claim of identical training dynamics.

Legacy Stage 3 backend. Native PyTorch CUDA in fp32 is the designated forward and autograd condition for the legacy Stage 3 equivalence and mechanism audit. The historical record did not fully pin the PyTorch, CUDA, and driver versions, so this names the retained execution path rather than a version-complete reproducible environment. The custom Stage 3 D64-bf16 backward path is disabled after failing its oracle and is not used for any ImageNet result. This path is distinct from the DiT-specific implementation discussed in Appendix S. Table C.1 reports the representative forward tests, backend capabilities, and tie audit without printing every synthetic tensor case.

Forward-equivalence result. Centered attention, attention outputs, and predictions agree within the predefined tolerances for all four representatives on the native fp32 path. The tie-sensitive subgradients above limit this result to forward behavior, not identical training dynamics.

Table C.1: Numerical Direct–intersection equivalence and gradient-semantics audit for the legacy Stage 3 ImageNet implementation. Forward comparisons cover the supported dimensions and precisions, with native CUDA fp32 designated as the recorded condition for this audit; centered attention and predictions pass, while exact ties retain distinct valid subgradients and the failing Stage 3 custom backward condition contributes no ImageNet evidence.

Test	Setting	Attention max error	Output max error	Prediction agreement	Backward result	Use
fp32 Direct representative	min	2.38419e-07	7.15256e-07	1	pass	forward-equivalent
fp32 Direct representative	stable_min	0	0	1	pass	forward-equivalent
fp32 Direct representative	l1	2.38419e-07	7.15256e-07	1	pass	forward-equivalent
fp32 Direct representative	l1_retrieval	2.38419e-07	8.34465e-07	1	pass	forward-equivalent
backend matrix	Stage3_ custom_ cuda/ bfloat16; D=32,64	0.000708207	–	–	fail	forward pass; appendix control
backend matrix	Stage3_ custom_ cuda/float32; D=32,64	–	–	–	unavailable	forward unavailable; appendix control
backend matrix	torch_cpu/ bfloat16; D=32,64	0.000750601	–	–	pass	forward pass; appendix control
backend matrix	torch_cpu/ float32; D=32,64	0	–	–	pass	forward pass; appendix control
backend matrix	torch_cuda/ bfloat16; D=32,64	0.000750601	–	–	pass	forward pass; appendix control
backend matrix	torch_cuda/ float32; D=32,64	0	–	–	pass	legacy designated condition
gradient semantics	exact q=k ties	–	–	100% in fp32	raw fail; tie-broken pass; 15111 ties	blocks training-dynamics equivalence only

D LEGACY-CHECKPOINT VALIDATION AND STRUCTURE

Preprocessing sensitivity of the legacy pair. The earlier ImageNet conversion used a generic evaluation transform—bilinear resize to 256, center crop to 224, and ImageNet mean/standard deviation—rather than the data configuration attached to the named timm checkpoint. Under that transform, the frozen Dot and legacy Direct checkpoints obtain 74.634% and 74.512% top-1 accuracy. Under the checkpoint's official configuration (bicubic interpolation, 0.9 center-crop fraction, and mean and standard deviation of 0.5), the same checkpoints obtain 81.392% and 79.318%, respectively. The official-transform Direct-minus-Dot difference is -2.074 percentage points, with paired 95% interval $[-2.292, -1.860]$. These are historical fixed-checkpoint measurements. The primary Q/K conversion in Table 1 and Appendix A uses a separately trained Direct endpoint and the official preprocessing.

Generic-transform sensitivity result. On all 50000 ImageNet validation images evaluated with the legacy generic transform, the Direct-minus-Dot point estimate is -0.122 percentage points, with paired 95% interval $[-0.334, 0.094]$ and $q = 0.26373$. The descriptive paired interval crosses zero. Conditional on this previously selected fixed checkpoint pair, it establishes neither equality, parity, nor non-inferiority. This result is a preprocessing sensitivity analysis, not the main ImageNet result.

Legacy selection and provenance. The legacy Direct checkpoint was selected under the generic transform. The retained bundle establishes its Q/K-only structure but does not fully archive the objective, optimizer schedule, or formal checkpoint-selection rule. Accordingly, this section audits the fixed endpoint rather than reconstructing a reproducible training protocol.

Bias-preserved Q/K-only structure. The endpoint-integrity audit compares each interpolation-path endpoint with its corresponding frozen checkpoint. Both comparisons yield 100% prediction agreement and zero maximum logit error (Table H.3), verifying reproduction of each checkpoint on the audited population. These are within-model checks, not agreement between Dot and Direct. The Direct checkpoint preserves every fused-QKV bias tensor and differs from Dot only in its Q/K weight rows: the state-dictionary audit reports all 12 Q/K tensors changed with all V tensors, QKV biases, and shared non-QKV tensors identical, so the Q/K-only intervention is clean at this legacy checkpoint.

Table D.1: Legacy generic-transform ImageNet-1K validation and checkpoint-structure audit. The table reports paired accuracy and logit effects on all 50,000 validation images, exact endpoint matching, layerwise tensor differences, and Q/K/V bias-restoration counterfactuals. This legacy result is retained only as a descriptive preprocessing-sensitivity analysis; no parity or non-inferiority conclusion is drawn from it. The Direct checkpoint preserves the Dot checkpoint's fused-QKV biases and differs only in Q/K weight rows, so the Q/K-only intervention is clean at this checkpoint.

Block	Condition	Primary estimate	Secondary estimate	95% CI / BH q	Decision
full validation	Dot checkpoint	top-1 74.634%	top-5 92.674%	–	frozen Dot model
full validation	Direct checkpoint	top-1 74.512%	top-5 92.492%	–	frozen Direct model
full validation	Direct - Dot accuracy	-0.122 pp	descriptive comparison	[-0.334, 0.094] pp; q=0.2637	paired interval includes zero; descriptive only
full validation	prediction agreement	89.618%	–	–	descriptive
full validation	D_KL(Dot ‖ Direct)	0.0901	–	–	descriptive
full validation	logit cosine	0.9765	–	–	descriptive
checkpoint structure	Q/K weights changed	yes	–	–	required
checkpoint structure	V weights identical	yes	–	–	required
checkpoint structure	shared non-QKV tensors identical	yes	–	–	required
checkpoint structure	missing fused-QKV bias tensors	0	–	–	clean QK-only intervention
bias counterfactual	restore_q	accuracy change -0.00 pp	KL from actual 0.0000	–	within threshold
bias counterfactual	restore_k	accuracy change -0.00 pp	KL from actual 0.0000	–	within threshold
bias counterfactual	restore_v	accuracy change -0.00 pp	KL from actual 0.0000	–	within threshold
bias counterfactual	restore_qk	accuracy change -0.00 pp	KL from actual 0.0000	–	within threshold
bias counterfactual	restore_qkv	accuracy change -0.00 pp	KL from actual 0.0000	–	within threshold

E CIFAR-100 OPERATOR SEARCH

Exploratory protocol. The ten-implementation comparison uses a 9.5M-parameter ViT-Tiny on CIFAR-100. Because it is a single-seed search reported at its best validation checkpoint, we treat it as exploratory operator-selection evidence rather than confirmatory accuracy evidence. This search was the sole purpose of the CIFAR-100 experiments; it discovered the Direct operator $-\max(0, q-k)$ (negated ReLU-subtraction), and no further CIFAR-100 results are reported in this paper.

Table E.1: Exploratory ViT-Tiny score-operator search on CIFAR-100, consolidated by exact row-softmax equivalence class. "Sep.?" marks scores that decompose into query-only and key-only terms and therefore yield query-independent normalized attention. Top-1 is single-seed best-validation accuracy, so the values are operator-selection evidence rather than confidence-bounded performance comparisons. Direct and histogram intersection occupy one forward-equivalent class; the incompatible historical $\min$ training trace remains an unresolved training-semantics diagnostic in Appendix L.

Score class	Property	Sep.?	Top-1
qk	dot baseline	No	60.58%
$-\mathrm{ReLU}(q-k) \equiv \min(q,k)$	Direct/intersection	No	60.31%
$\lvert q-k\rvert$	distance	No	11.65%
$(q-k)^2$	distance	No	9.18%
$\max(q,k)$	symmetric comparison	No	5.79%
$\mathrm{ReLU}(q+k)$	thresholded sum	No	5.79%
$q+k$	linear	Yes	5.69%
$\mathrm{ReLU}(q-k)$	wrong-sign Direct	No	5.46%
$q-k$	linear	Yes	5.26%

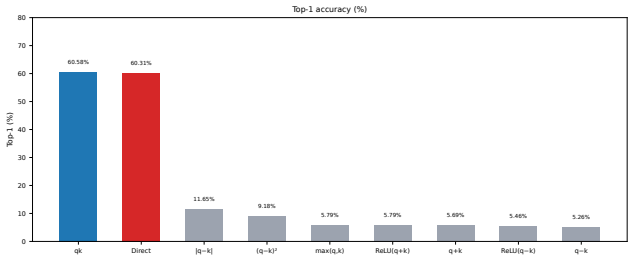


Figure E.1: Single-seed best-validation top-1 accuracy per score class from the exploratory CIFAR-100 search (Table E.1); the dot baseline (blue) and Direct/intersection (red) are highlighted; top-1 is higher is better (↑).

Recorded operator outcomes. Table E.1 shows that the recorded dot and Direct representatives both exceed 60% top-1 accuracy, while seven other non-equivalent score classes stay below 12%. Replacing multiplication is therefore not trivial under this protocol. The historical 500-step run labeled $\min(q,k)$ reached 7.33%, but Eq. (4) makes that trajectory incompatible with treating the formulas as distinct operators. Gate 0 subsequently verified their forward attention and predictions on the audited implementation and native backend. The retained historical record is insufficient to identify why that short training trajectory diverged, so 7.33% remains provenance for an unresolved training diagnostic and is never used to argue that Direct is uniquely successful.

Separability explains only part of the search. The separability theorem correctly predicts the failure of $q+k$ and $q-k$, whose query-only terms vanish under row-wise softmax. Non-separability alone is insufficient: $|q-k|$, $(q-k)^2$, $\max(0, q+k)$, and $\max(0, q-k)$ are non-separable but perform poorly. The Direct–intersection class combines non-separability with a sign and threshold geometry that induce a useful key ordering in the recorded run; raw directional asymmetry is not a distinguishing explanation.

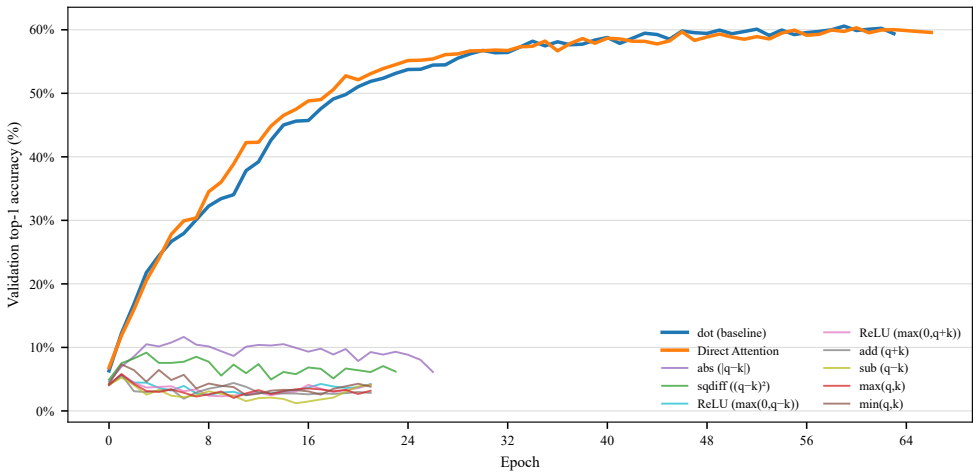


Figure E.2: Exploratory single-seed ViT-Tiny/CIFAR-100 best-validation trajectories for the ten recorded score formulas. Curves show validation top-1 accuracy over the shared training budget and are grouped in the text by row-softmax equivalence rather than by raw formula label. The Direct and Dot runs exceed 60%, while the other recorded classes remain below 12%; these single-seed outcomes select candidates but provide no confidence-bounded performance comparison. The incompatible historical minimum-formula trace remains causally unresolved and is not an independent operator result.

Score sign controls key ordering. The comparison between $\max(0, q - k)$ and $-\max(0, q - k)$ is especially informative. The two recorded single-seed arms differ only in sign; this sign reverses the key ordering rewarded by softmax, and their recorded best accuracies differ by 54.85 percentage points. The comparison records this co-occurrence but does not isolate the sign as the cause of the performance gap. Complete trajectories and the separate gradient/tie audit appear here and in Appendix L. Together, the search selects the Direct–intersection class for subsequent scaling, but it does not establish seed-robust superiority, a shared mechanism, or training-dynamics equivalence.

F LEGACY-CHECKPOINT RETRIEVAL GEOMETRY AND CROSSINGS

All numerical retrieval and intervention results in this section and Appendices G–I use the legacy pair identified in Appendix D. They characterize that pair's inference behavior, not the primary conversion endpoint.

Table F.1: Predefined inference-only mechanism tests for the frozen legacy Dot and Direct checkpoints (summarized in Section 5). Shared geometry requires at least 10 of 12 layers to pass centered-score Spearman, high-margin agreement, and top-5-overlap thresholds simultaneously; the threshold representation requires attention cosine at least 0.99 and top-5 overlap at least 0.95; the coupled Q/K–score path requires an accuracy barrier no larger than 1 percentage point and $D_{\mathrm{KL}}(p_{\mathrm{Dot}} \| p_{\mathrm{path}}) \leq 0.1$; the component intervention requires the full score to beat all three controls with paired 95% intervals above zero and BH-adjusted $q < .05$. Geometry and components pass while threshold ranking and path continuity fail, supporting substantial shared structure while rejecting functional reparameterization.

Test	Estimate and criterion	Decision
Shared geometry	11/12 layers; require 10/12	Pass
Threshold features	cosine 0.9951; top-5 0.7792; require $0.99/0.95$	Fail
Q/K–operator path	barrier 12.60 pp and KL 0.712; require ≤ 1 pp and ≤ 0.1	Fail
Score components	3/3 predefined controls; $q_{\max} = 8.755 \times 10^{-145}$	Pass

Layerwise geometry protocol. For each transformer layer, the audit compares centered scores, high-margin key preferences, top-5 keys, attention divergence, and crossing density. Figure F.1 gives the layerwise profile and Table F.2 reports all 12 layer aggregates. Head-level records are summarized at the layer level because the claim gate is defined per layer.

Table F.2: Complete 12-layer retrieval-geometry summary on the fixed 5,000-image population. Centered-score Spearman correlation, high-margin key-preference agreement, top-5 overlap, attention divergence, and crossing density are aggregated by layer. Eleven of 12 layers meet all three predefined alignment thresholds, above the required 10, so the shared-geometry criterion is met for this legacy checkpoint pair.

Layer	Centered Spearman	High-margin agreement	Top-5 overlap	Attention KL	Crossing density	All three pass
1	0.841647	0.976055	0.64923	0.162483	0.156485	yes
2	0.905967	0.994239	0.71385	0.130641	0.122347	yes
3	0.903687	0.993646	0.738479	0.130602	0.12407	yes
4	0.877667	0.990482	0.647702	0.136495	0.143981	yes
5	0.902582	0.995692	0.696784	0.119093	0.128778	yes
6	0.889587	0.992261	0.645958	0.142105	0.136055	yes
7	0.888477	0.990344	0.640857	0.161088	0.136146	yes
8	0.864927	0.982729	0.619815	0.233621	0.148432	yes
9	0.820054	0.970299	0.611639	0.285723	0.173657	yes
10	0.782208	0.960554	0.619881	0.338407	0.19441	yes
11	0.738619	0.947073	0.599946	0.448562	0.216621	yes
12	0.666209	0.915566	0.51937	0.616208	0.249227	no

Shared retrieval geometry meets the alignment gate. Eleven of twelve layers jointly satisfy the centered-score Spearman, high-margin preference-agreement, and top-5-overlap thresholds; the final layer (block 11) misses the centered-score criterion, and attention KL and crossing density rise descriptively in the later layers. The predefined gate requires at least ten of twelve, so it passes. The checkpoints therefore show substantial layerwise alignment, while threshold-level key ranking, connectivity, and spatial evidence still differ (see Appendices G and H).

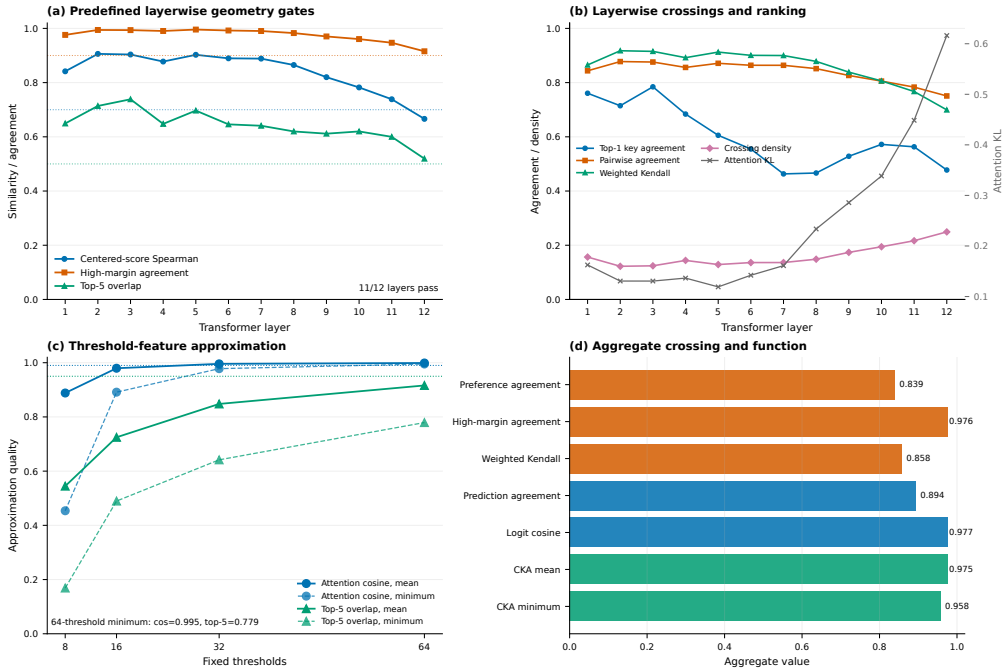


Figure F.1: Complete layerwise retrieval-geometry audit for the frozen ViT-Small/ImageNet-1K Dot and Direct checkpoints on 5,000 fixed images. The panels report centered-score Spearman correlation, high-margin preference agreement, top-5 key overlap, attention divergence, and crossing density, with layers as the statistical unit. A layer passes only when the first three metrics meet their predefined thresholds of 0.70, 0.90, and 0.50. 11 of 12 layers pass, meeting the required 10/12, so the shared-retrieval-geometry gate passes; top-5 key overlap remains the weakest criterion and declines toward the output.

G Threshold Representation and Symmetry Probes

Threshold-feature approximation. For bounded coordinates $q, k \in [L, U]$, the coordinatewise minimum has the threshold-feature representation

$$\min(q, k) = L + \int_L^U \mathbf{1}[q \geq t]\, \mathbf{1}[k \geq t]\, dt. \tag{G.1}$$

The integral is approximated with fixed grids of 8, 16, 32, and 64 thresholds; bounds are selected on the intervention population and evaluated on the disjoint geometry population. At 64 thresholds the minimum attention cosine is 0.9951, while minimum top-5 overlap 0.7792 and out-of-bound rate 1.069e-05 expose the remaining approximation error.

Symmetry scope. The same table reports simultaneous Q/K translation, orthogonal rotation, and coordinate-permutation probes. These diagnose score symmetries; they do not imply that two trained checkpoints occupy the same representation.

Table G.1: Threshold-feature approximation and symmetry probes on disjoint calibration and evaluation populations. Rows report attention cosine, top-5 overlap, and out-of-bound error for 8–64 thresholds, followed by translation, rotation, and coordinate-permutation invariance tests. The 64-threshold cosine criterion passes but top-5 overlap fails, so high similarity does not preserve key ranking.

Block	Condition	Primary metric	Secondary metric	Maximum error	Interpretation
threshold features	8 thresholds	cosine mean/ min 0.8882/ 0.4538	top-5 mean/ min 0.5446/ 0.1688	OOB 1.069e-05; attn 0.6958	approximation diagnostic
threshold features	16 thresholds	cosine mean/ min 0.9796/ 0.8913	top-5 mean/ min 0.7248/ 0.4896	OOB 1.069e-05; attn 0.3717	approximation diagnostic
threshold features	32 thresholds	cosine mean/ min 0.9958/ 0.9780	top-5 mean/ min 0.8479/ 0.6417	OOB 1.069e-05; attn 0.2001	approximation diagnostic
threshold features	64 thresholds	cosine mean/ min 0.9989/ 0.9951	top-5 mean/ min 0.9161/ 0.7792	OOB 1.069e-05; attn 0.0798	64-threshold predefined gate
symmetry	direct: direct_ permutation	expected invariant: yes	max centered-score error 2.289e-05	attention 1.490e-07	invariance probe
symmetry	direct: direct_ rotation	expected invariant: no	max centered-score error 3.516e+01	attention 8.187e-01	non-invariance control
symmetry	direct: direct_ translation	expected invariant: yes	max centered-score error 1.526e-05	attention 1.192e-07	invariance probe
symmetry	dot: dot_ permutation	expected invariant: yes	max centered-score error 3.052e-05	attention 2.682e-07	invariance probe
symmetry	dot: dot_ rotation	expected invariant: yes	max centered-score error 7.629e-05	attention 4.172e-07	invariance probe
symmetry	dot: dot_ translation	expected invariant: no	max centered-score error 5.162e+01	attention 5.732e-01	non-invariance control

Threshold features preserve similarity but not key ranking. At 64 thresholds, the minimum attention cosine of 0.9951 exceeds the predefined 0.99 criterion, but the minimum top-5 overlap of 0.7792 remains below 0.95, so the threshold-representation gate fails despite the low out-of-bound rate. The symmetry probes recover the operators' distinct expected invariances: Direct preserves simultaneous translation and coordinate permutation but not orthogonal rotation, whereas dot product preserves rotation and permutation but not simultaneous translation. These results support a bounded threshold-feature interpretation and operator-specific symmetry characterization, but not exact finite-grid equivalence, preserved key ranking, a shared trained representation, or functional reparameterization.

H LEGACY-CHECKPOINT Q/K–OPERATOR CONNECTIVITY AND SCORE COMPONENTS

Connectivity result. The path audit couples interpolated Q/K parameters and the score operator. Exact Dot and Direct endpoints match their corresponding frozen checkpoint outputs, but the tested diagonal path has a 12.60-pp accuracy barrier and maximum $D_{\mathrm{KL}}(p_{\mathrm{Dot}} \| p_{\mathrm{path}})$ of 0.712. This rejects the predefined shared-functional-basin criterion.

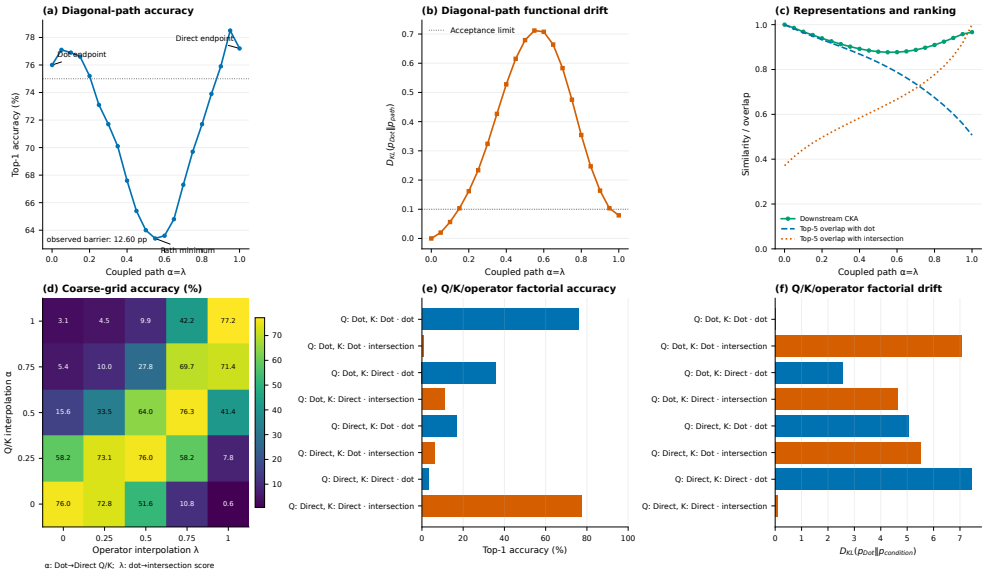


Figure H.1: Complete Q/K–operator connectivity audit on the fixed 1,000-image intervention population. The coarse phase diagram spans Dot-to-Direct Q/K interpolation on one axis and dot-to-intersection score interpolation on the other; the diagonal panels report top-1 accuracy, Dot-to-path logit KL, downstream CKA, and top-key overlap across 21 coupled points. The factorial panels retain all eight combinations of Dot/Direct Q source, Dot/Direct K source, and dot/intersection score, together with exact endpoints and the restored Dot V-bias control. Endpoint outputs match their frozen checkpoints, but the 12.60-point accuracy barrier and maximum KL 0.712 fail the predefined connectivity gate.

Component-factorial result. The score-component interventions compare the full L1-plus-coordinate-bias score with three predefined controls and three supporting substitutions. Only distance-only, bias-only, and shuffled coordinate bias enter the acceptance test. All three have full-minus-control intervals above zero after correction; the substitutions diagnose checkpoint-specific sensitivity and are not additional acceptance tests.

Table H.1: Representative points from the 21-point coupled Q/K–operator path on 1,000 images. The 12.60-point accuracy barrier and maximum Dot-to-path KL 0.712 fail the predefined connectivity limits.

Condition	Accuracy	$D_{\mathrm{KL}}(\mathrm{Dot} \parallel \mathrm{path})$	Role
$\alpha = \lambda = 0.00$	76.0%	0.000	endpoint
$\alpha = \lambda = 0.25$	73.1%	0.234	interior
$\alpha = \lambda = 0.50$	64.0%	0.679	interior
$\alpha = \lambda = 0.55$	63.4%	0.712	minimum / barrier peak
$\alpha = \lambda = 0.75$	69.7%	0.475	interior
$\alpha = \lambda = 1.00$	77.2%	0.079	endpoint

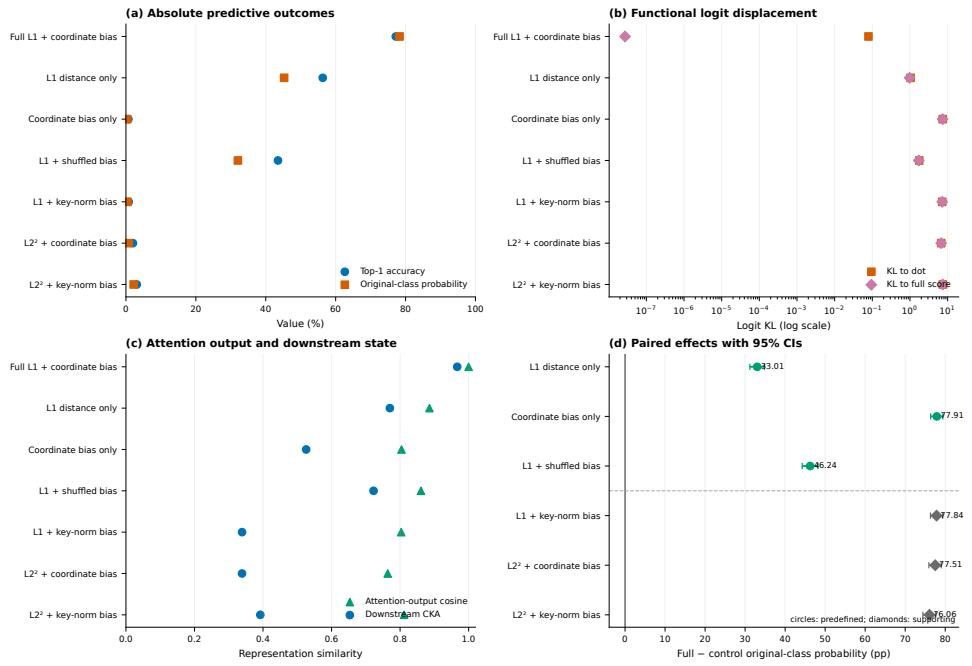


Figure H.2: Score-component interventions on the same 1,000-image population. The paired forest reports full-score minus control original-class probability in percentage points with 95% bootstrap intervals; the first three rows are predefined controls and the remaining three are supporting substitutions. The absolute panels report accuracy, class probability, Dot-reference KL, attention-output cosine, and downstream CKA for all seven score conditions. All three predefined contrasts pass after correction, showing checkpoint-level dependence on the combined distance and coordinate-bias terms, but this result does not repair the failed Q/K–operator path.

Table H.2: Complete two-by-two-by-two factorial over Dot/Direct Q source, K source, and score operator on the same 1,000 images.

Q source	K source	Score	Accuracy	Dot-ref. KL	CKA
Dot	Dot	dot	76.0%	0.000	1.000
Dot	Dot	intersection	0.6%	7.070	0.575
Dot	Direct	dot	35.9%	2.558	0.461
Dot	Direct	intersection	11.1%	4.633	0.549
Direct	Dot	dot	16.7%	5.042	0.595
Direct	Dot	intersection	6.2%	5.497	0.642
Direct	Direct	dot	3.1%	7.440	0.392
Direct	Direct	intersection	77.2%	0.079	0.967

Component dependence does not repair path discontinuity. The full Direct score outperforms all three predefined component controls with corrected paired intervals above zero, so this checkpoint's inference behavior depends on the combined L1-distance and coordinate-sum-bias construction rather than either isolated or shuffled component. However, exact endpoint matching coexists with a 12.60-pp accuracy barrier and maximum $D_{\mathrm{KL}}(p_{\mathrm{Dot}} \| p_{\mathrm{path}})$ of 0.712 along the tested coupled path. The V-bias control rules out V-bias differences as a confound in the endpoint comparison. These results support checkpoint-specific component dependence at the endpoints while rejecting shared-functional-basin and functional-reparameterization interpretations.

Table H.3: Frozen-checkpoint endpoint integrity and the restored Dot V-bias control. Both endpoints match exactly; the structural control retains 77.2% accuracy and Dot-reference KL 0.079.

Block	Condition	Estimates	Decision / role
exact endpoint	direct_endpoint	logit max error 0.0e+00; agreement 100.0%	pass
exact endpoint	dot_endpoint	logit max error 0.0e+00; agreement 100.0%	pass
structural control	Direct endpoint with restored Dot V bias	accuracy 77.2%; D_{KL}(Dot ‖ path) 0.079	bias confound control

Table H.4: Score-component interventions on 1,000 images. Absolute accuracy, original-class probability, Dot-reference KL, attention-output cosine, and downstream CKA are reported for the full score and six substitutions; paired full-minus-control intervals distinguish three predefined controls from three supporting substitutions. All three predefined component contrasts pass, but that checkpoint-specific dependence does not repair path discontinuity.

Block	Condition	Primary estimate	Secondary estimate	95% CI / BH q	Decision / role
score components	full_l1_ coordinate_ bias	accuracy 77.2%; class prob. 78.3%	KL 0.079; CKA 0.967	–	full score
score components	l1_distance_ only	accuracy 56.3%; class prob. 45.3%	KL 1.034; CKA 0.770	full-control [31.19, 34.82] pp; q=8.76e-145	predefined gate control
score components	coordinate_ sum_bias_only	accuracy 0.7%; class prob. 0.4%	KL 7.385; CKA 0.526	full-control [76.37, 79.40] pp; q=5.48e-165	predefined gate control
score components	l1_shuffled_ bias	accuracy 43.5%; class prob. 32.0%	KL 1.760; CKA 0.723	full-control [44.26, 48.22] pp; q=4.51e-157	predefined gate control
score components	l1_key_norm_ bias	accuracy 0.8%; class prob. 0.4%	KL 7.175; CKA 0.339	full-control [76.31, 79.37] pp; q=5.48e-165	supporting substitution
score components	l2_coordinate_ bias	accuracy 2.0%; class prob. 0.8%	KL 6.754; CKA 0.339	full-control [75.93, 79.00] pp; q=5.48e-165	supporting substitution
score components	full_l2_key_ norm_bias	accuracy 3.1%; class prob. 2.2%	KL 7.440; CKA 0.392	full-control [74.42, 77.62] pp; q=5.48e-165	supporting substitution

I LEGACY-CHECKPOINT PATCH-REMOVAL INTERVENTIONS

Model-specific removal protocol. Each checkpoint is evaluated with its own saliency map. Top-, bottom-, and random-ranked patches are removed at 10%, 25%, and 50% on the disjoint 1000-image population. The resulting 18 aggregate conditions and paired gate summaries are reported compactly.

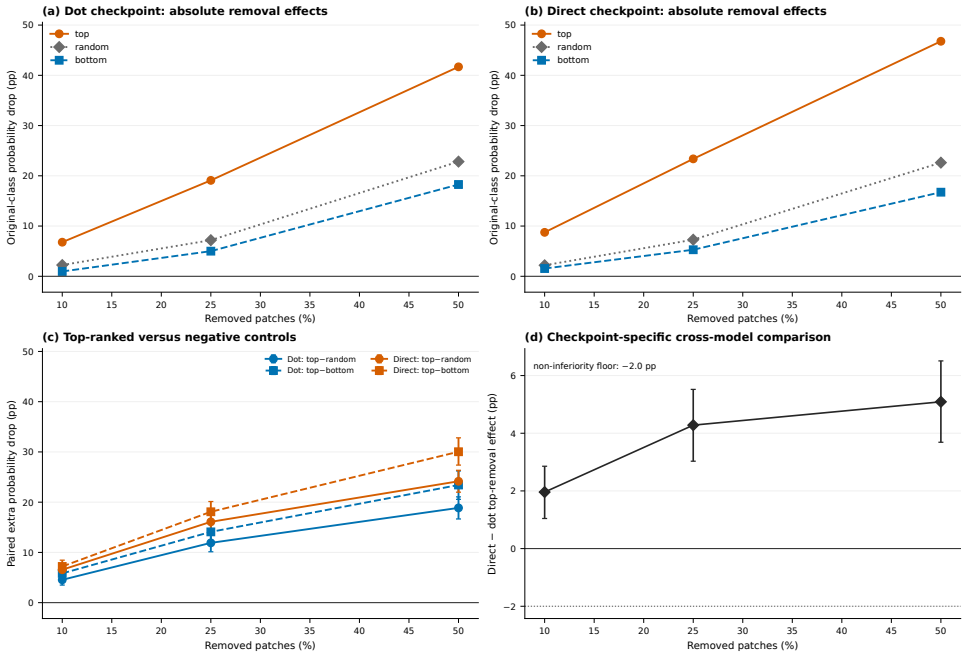


Figure I.1: Model-specific patch removal on the disjoint 1,000-image population. Each checkpoint uses its own saliency ranking, and patches are removed from the top, random, or bottom 10%, 25%, and 50% of positions. Points show mean original-class probability drop and bars show paired 95% bootstrap intervals; larger positive drops indicate greater predictive consequence. All top-minus-random and top-minus-bottom gates pass for both checkpoints, which supports within-model faithfulness but not shared salient regions.

Patch removal supports within-model faithfulness only. Using each checkpoint's own saliency map, removing top-ranked patches causes a larger original-class probability drop than removing random- or bottom-ranked patches at all three removal budgets for both models, so all predefined within-model comparisons pass. Because the ranking is recomputed separately for each checkpoint, the larger Direct drops are not a shared-intervention comparison and do not show that the two models select the same regions. The result supports the faithfulness of each model's own patch ranking, but it does not establish a shared retrieval mechanism, functional reparameterization, or mechanism-level superiority of Direct attention.

Table I.1: Absolute model-specific patch-removal outcomes on 1,000 disjoint images. These aggregate estimates cover Dot and Direct own-saliency top, random, and bottom removal at 10, 25, and 50 percent; larger original-class probability drop indicates greater predictive consequence.

Model	Removed	Ranking	Original-class probability drop
direct	10%	bottom	0.0154746
direct	25%	bottom	0.0528417
direct	50%	bottom	0.167393
direct	10%	random	0.0219224
direct	25%	random	0.07284
direct	50%	random	0.22618
direct	10%	top	0.0874638
direct	25%	top	0.233717
direct	50%	top	0.467767
dot	10%	bottom	0.00963299
dot	25%	bottom	0.0500329
dot	50%	bottom	0.182674
dot	10%	random	0.0224072
dot	25%	random	0.0718965
dot	50%	random	0.22825
dot	10%	top	0.0678676
dot	25%	top	0.190914
dot	50%	top	0.416866

Table I.2: Paired patch-removal gate effects and cross-model comparison. Top-minus-random and top-minus-bottom intervals test each checkpoint's own ranking at every removal budget, while Direct-minus-Dot effects compare magnitude. Every row is a paired gate summary and passes; the last column gives its maximum BH-adjusted q-value. The results support own-map faithfulness rather than shared salient regions.

Model	Removed	Paired probability-drop effects	Max BH q
dot	10%	top-random: 4.5 [3.5,5.6] pp; top-bottom: 5.8 [4.6,7.0] pp	3.77e-14
dot	25%	top-random: 11.9 [10.1,13.7] pp; top-bottom: 14.1 [12.1,16.2] pp	1.07e-29
dot	50%	top-random: 18.9 [16.7,21.1] pp; top-bottom: 23.4 [20.6,26.2] pp	8.03e-49
direct	10%	top-random: 6.6 [5.4,7.7] pp; top-bottom: 7.2 [6.0,8.5] pp; Direct-dot: 2.0 [1.0,2.9] pp	1.01e-06
direct	25%	top-random: 16.1 [14.4,17.8] pp; top-bottom: 18.1 [16.0,20.1] pp; Direct-dot: 4.3 [3.0,5.5] pp	1.61e-13
direct	50%	top-random: 24.2 [22.0,26.3] pp; top-bottom: 30.0 [27.4,32.8] pp; Direct-dot: 5.1 [3.7,6.5] pp	3.16e-11

J LEGACY-CHECKPOINT IMAGENET-S LOCALIZATION AND EXPLAINABILITY

Every localization metric, explanation map, and control in this section uses the legacy checkpoint pair from Appendix D. These diagnostics have not been repeated for the primary conversion endpoint.

J.1 HISTORICAL LOCALIZATION AUDIT

Controlled localization result. Localization uses all 4580 valid ImageNet-S masks and also reports the 3504 jointly-correct subset. Original saliency is compared with spatial shuffle, random maps, entropy-preserving permutation, and a center prior. Foreground energy, pixel average precision, pointing-game accuracy, and IoU are retained. Although Direct remains close to dot attention, both models fail the foreground-energy and pixel-AP center-prior contrasts; the localization gate therefore fails.

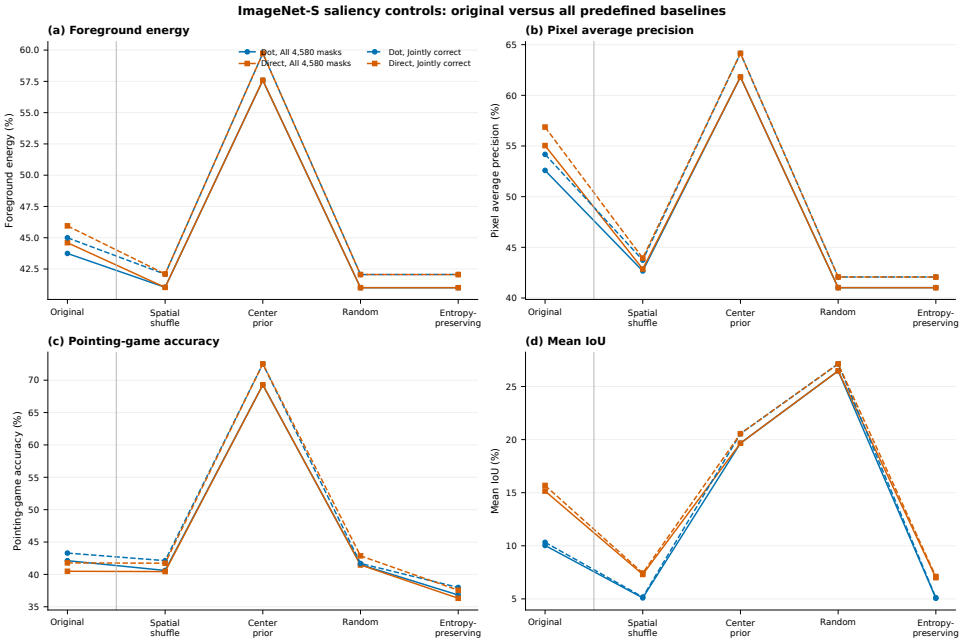


Figure J.1: ImageNet-S localization audit on all 4580 valid masks, with the 3504-image jointly-correct subset reported separately. Foreground energy, pixel average precision, pointing game, and intersection over union are evaluated for original maps and random, spatially shuffled, entropy-preserving, and center-prior controls. Paired effects are oriented as original minus control, so positive values favor the learned map. Both checkpoints fail the required center-prior contrasts for foreground energy and pixel average precision; favorable Direct-minus-Dot differences therefore remain descriptive rather than mechanistic evidence.

Historical explainer scope. The historical rollout, gradient-weighted rollout, ViT Grad-CAM, integrated gradients, patch occlusion, and Transformer Input Sampling diagnostics below (Abnar & Zuidema, 2020; Selvaraju et al., 2017; Sundararajan et al., 2017; Englebert et al., 2023) are retained as historical diagnostics. The common-population audit below supersedes those eight-image measurements within this legacy-checkpoint study.

Center-prior controls reject the localization claim. Across all 4,580 valid masks, Direct exceeds dot by 0.86 percentage points in foreground energy, 2.45 points in pixel AP, and 5.12 points in mean IoU, with paired intervals wholly above zero, while pointing game favors dot by 1.64 points. Nevertheless, each checkpoint's original map is worse than the center prior on the primary metrics: the original-minus-prior effects are -12.99 and -6.77 points for Direct and -13.84 and -9.22 points for dot in foreground energy and pixel AP, respectively. The predefined localization gate

Table J.1: ImageNet-S localization aggregates on all 4,580 valid masks. Original Dot and Direct saliency is listed alongside random, shuffled, entropy-preserving, and center-prior controls under foreground energy, pixel average precision, pointing game, and intersection over union. Paired contrasts appear in Table J.2.

Block	Condition	Foreground energy	Pixel AP	Other result	95% CI / BH q	Role
localization aggregate	direct/ center_prior	0.575871	0.618094	pointing 0.693; IoU 0.197	–	control
localization aggregate	direct/ entropy_ preserving_ permutation	0.410032	0.410126	pointing 0.363; IoU 0.070	–	control
localization aggregate	direct/ original	0.446015	0.550392	pointing 0.405; IoU 0.152	–	original
localization aggregate	direct/ random	0.410064	0.410196	pointing 0.415; IoU 0.265	–	control
localization aggregate	direct/ spatial_ shuffle	0.410303	0.428625	pointing 0.404; IoU 0.073	–	control
localization aggregate	dot/center_ prior	0.575871	0.618094	pointing 0.693; IoU 0.197	–	control
localization aggregate	dot/entropy_ preserving_ permutation	0.410079	0.41023	pointing 0.368; IoU 0.051	–	control
localization aggregate	dot/original	0.437441	0.525917	pointing 0.421; IoU 0.100	–	original
localization aggregate	dot/random	0.410063	0.410191	pointing 0.415; IoU 0.265	–	control
localization aggregate	dot/spatial_ shuffle	0.410379	0.426655	pointing 0.406; IoU 0.051	–	control

therefore fails, and the older eight-image diagnostics remain provenance rather than claim-bearing evidence.

J.2 Common-Population Explainability Protocol

Declared method population. The inference-only audit declares 17 dense spatial methods before inspection. A common set of 1,000 jointly-correct images supports every cross-method comparison; raw attention, rollout, gradient rollout, LeGrad, Grad-CAM, and Input×Gradient additionally use all 4,580 valid masks. Localization targets the ground-truth class, which equals the original prediction on the common set. Faithfulness uses each checkpoint's original class probability. Concept-level and local-surrogate explainers are excluded because they do not produce like-for-like ViT patch maps.

J.3 Control-Qualified Quantitative Comparison

Qualified effects are method- and metric-specific. Several attention-native and Transformer-specific methods produce positive Direct-minus-dot localization effects, but their localization-control pass counts range from zero to three of four. The CAM family contains substantial dot-favored

Table J.2: Paired ImageNet-S localization controls for the legacy Dot and Direct checkpoints. Positive intervals against shuffled, random, and entropy-preserving controls pass, whereas both center-prior comparisons fail; the localization evidence is therefore descriptive.

Block	Condition	Foreground energy	Pixel AP	Other result	95% CI / BH q	Role
paired control	dot_original_minus_spatial_shuffle	0.027062	0.0992622	–	FE [0.024, 0.030], q=3.4e-69; AP [0.092, 0.106], q=5.49e-111	pass
paired control	dot_original_minus_center_prior	-0.13843	-0.0921769	–	FE [-0.143, -0.133], q=0; AP [-0.100, -0.085], q=5.31e-146	fail
paired control	dot_original_minus_random	0.0273784	0.115726	–	FE [0.025, 0.030], q=2.53e-79; AP [0.109, 0.122], q=2.79e-161	pass
paired control	dot_original_minus_entropy_preserving_permutation	0.0273622	0.115687	–	FE [0.025, 0.030], q=2.9e-79; AP [0.109, 0.122], q=2.79e-161	pass
paired control	direct_original_minus_spatial_shuffle	0.0357126	0.121767	–	FE [0.033, 0.039], q=4.91e-117; AP [0.115, 0.129], q=7.58e-160	pass
paired control	direct_original_minus_center_prior	-0.129856	-0.067702	–	FE [-0.135, -0.125], q=0; AP [-0.075, -0.060], q=9.36e-90	fail
paired control	direct_original_minus_random	0.0359509	0.140196	–	FE [0.033, 0.039], q=6.27e-136; AP [0.133, 0.147], q=1.12e-219	pass
paired control	direct_original_minus_entropy_preserving_permutation	0.0359833	0.140266	–	FE [0.033, 0.039], q=5.31e-136; AP [0.133, 0.147], q=9.34e-220	pass

Table J.3: Historical supporting explainability diagnostics retained for context. Rows summarize the earlier rollout, gradient, CAM, integrated-gradient, occlusion, and sampling comparisons, including the superseded small-population evidence. These entries are non-authoritative and do not override the controlled Stage 3 localization failure.

Block	Condition	Foreground energy	Pixel AP	Other result	95% CI / BH q	Role
paired model comparison	Direct - dot original saliency	0.00857396	0.0244748	–	FE [0.008, 0.009], q=2.6e-89; AP [0.023, 0.026], q=2.35e-141	non-inferiority passes
supporting explainer	gradient rollout	–	–	Tiny/CIFAR cosine 0.846; ViT-S/ ImageNet cosine 0.820; Stage2/ ImageNet-S cosine 0.805	–	historical/ supporting only
supporting explainer	Grad-CAM	–	–	Tiny/CIFAR cosine 0.316; ViT-S/ ImageNet cosine 0.826; Stage2/ ImageNet-S cosine 0.894	–	historical/ supporting only
supporting explainer	integrated gradients	–	–	Tiny/CIFAR cosine 0.751; ViT-S/ ImageNet cosine 0.654	–	historical/ supporting only
supporting explainer	patch occlusion	–	–	Tiny/CIFAR cosine 0.793; ViT-S/ ImageNet cosine 0.932; Stage2/ ImageNet-S cosine 0.949	–	historical/ supporting only
supporting explainer	Transformer Input Sampling	–	–	Tiny/CIFAR cosine 0.947; ViT-S/ ImageNet cosine 0.960	–	historical/ supporting only

effects, while integrated gradients fails the parameter-randomization sanity criterion (Adebayo et al., 2018). The audit therefore supports only individual claim-eligible method–metric effects, not an across-method Direct advantage.

J.4 Metric Families and Negative Controls

No explainer family provides uniform evidence for Direct. Positive Direct effects on sign-oriented deletion AUC coexist with frequently negative insertion effects and localization effects that sometimes favor dot. Map agreement and transformation stability describe consistency, not

Table J.4: Predeclared dense spatial explainability suite. The table names each method family, class target, Transformer layer or perturbation parameter, relative cost, evaluation population, and retained constant-map count. Raw attention, rollout, gradient rollout, LeGrad, Grad-CAM, and Input-times-Gradient additionally cover all 4,580 valid ImageNet-S masks; every method uses the common 1,000-image population. Constant maps remain in all aggregates rather than being silently removed.

Method	Family	Target	Layer / parameter	Cost	Population	Constant maps
Raw CLS	attention	n/a	last block; mean heads	low	4,580 + common	0/2,000
Rollout	attention	n/a	all blocks + residual	low	4,580 + common	0/2,000
Attention flow	attention	n/a	all-block capacity flow	high	common 1,000	0/2,000
Gradient rollout	attention	ground-truth	all blocks; positive A×grad	medium	4,580 + common	0/2,000
Chefer	transformer	ground-truth	all-block Deep-Taylor/LRP	medium	common 1,000	0/2,000
LeGrad	transformer	ground-truth	all layer-wise feature targets	medium	4,580 + common	0/2,000
Grad-CAM	cam	ground-truth	last block norm1	medium	4,580 + common	0/2,000
Grad-CAM++	cam	ground-truth	last block norm1	medium	common 1,000	6/2,000
HiResCAM	cam	ground-truth	last block norm1	medium	common 1,000	0/2,000
XGradCAM	cam	ground-truth	last block norm1	medium	common 1,000	0/2,000
Saliency	input gradient	ground-truth	absolute input gradient	medium	common 1,000	0/2,000
Input × Gradient	input gradient	ground-truth	positive input × gradient	medium	4,580 + common	0/2,000
SmoothGrad	input gradient	ground-truth	25 noise samples	high	common 1,000	0/2,000
Integrated gradients	input gradient	ground-truth	32 steps; blur baseline	high	common 1,000	0/2,000
Patch occlusion	perturbation	ground-truth	16×16; blur baseline	high	common 1,000	0/2,000
TIS	perturbation	ground-truth	512 masks; keep=.5	high	common 1,000	0/2,000
RISE	perturbation	ground-truth	512 token masks; keep=.5	high	common 1,000	0/2,000

faithfulness, and the top-versus-random and top-versus-bottom controls pass unevenly across methods and removal budgets. Results must therefore remain stratified by method, metric family, and control status.

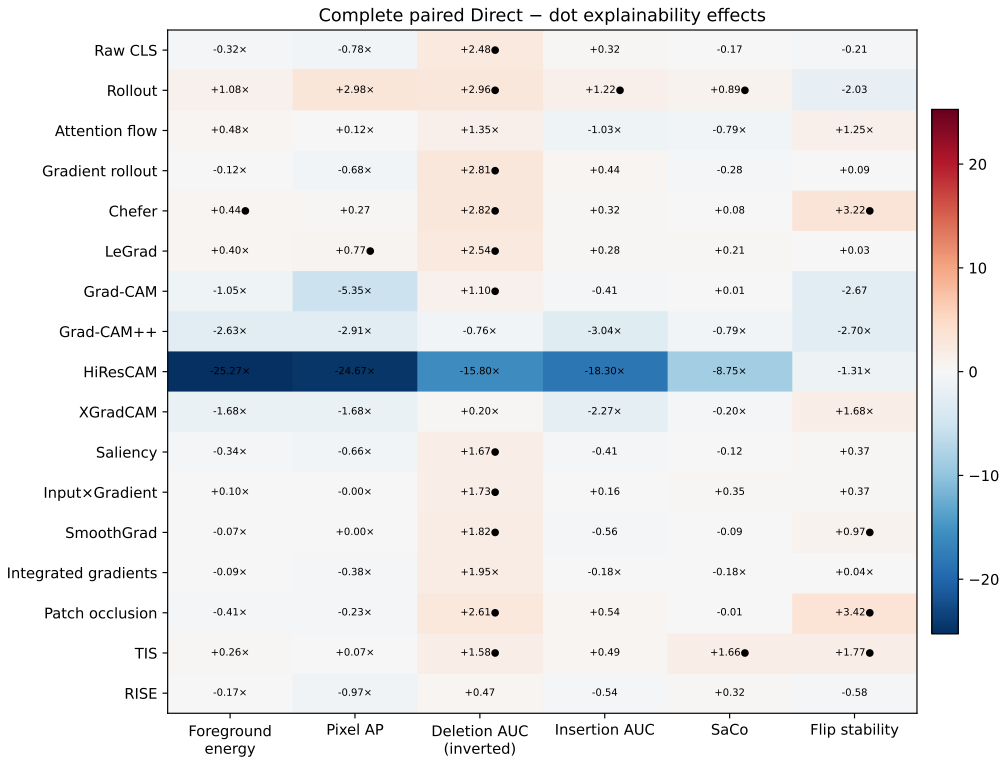


Figure J.2: Complete paired Direct-minus-Dot explainability effects for all 17 predeclared methods on the common 1,000-image, jointly-correct ImageNet-S population. Columns cover localization, faithfulness, and flip stability; deletion area under the curve is sign-reversed so positive consistently favors Direct. Filled circles denote paired 95% intervals in the favorable direction with within-family Benjamini–Hochberg-adjusted $q < .05$, while crosses identify a required sanity or control failure. Positive localization remains descriptive when random, shuffle, entropy-preserving, center-prior, or faithfulness qualification fails.

J.5 Attention and Transformer-Specific Explanations

Attention-native gains remain control-limited. Rollout favors Direct on the reported localization and faithfulness effects, but it passes none of the four localization controls. Attention flow shows insertion-AUC effects that favor dot, while Chefer and LeGrad pass only three of four spatial controls. These methods provide method-specific observations rather than evidence that dot and Direct share an attention mechanism.

J.6 Grad-CAM Family

CAM evidence often favors dot attention. Grad-CAM and Grad-CAM++ favor dot on foreground energy and pixel AP, while HiResCAM favors dot by 24.67 and 25.27 percentage points on those metrics and also favors dot on deletion, insertion, and SaCo. XGradCAM's localization intervals cross zero; although Grad-CAM deletion favors Direct, its insertion effect favors dot, and Grad-CAM++ contributes four retained constant maps. These mixed results rule out a general Direct advantage under CAM-based explanations.

J.7 Input-Gradient Explanations

Input-gradient effects are mixed and sanity-sensitive. Input×Gradient and saliency have localization intervals that cross zero, whereas SmoothGrad significantly favors dot on foreground energy and pixel AP. All four methods favor Direct on sign-oriented deletion, but insertion favors dot for saliency,

Table J.5: Paired Direct–Dot localization and sign-oriented deletion effects for attention-native methods: Attention flow, Rollout, Gradient rollout, Raw CLS, on the common 1,000-image ImageNet-S population. Effects are in percentage points with positive favoring Direct; intervals are paired 95% bootstrap intervals and q values use within-family Benjamini–Hochberg correction. Statistical significance alone does not qualify a claim when a required sanity or spatial control fails.

Method	Metric	Direct-dot	95% CI	BH q
Attention flow	Foreground energy	+0.48 pp	[+0.07, +0.90]	0.0376
Attention flow	Pixel AP	+0.12 pp	[-0.23, +0.45]	0.00744
Rollout	Foreground energy	+1.08 pp	[+0.90, +1.26]	8.75e-30
Rollout	Pixel AP	+2.98 pp	[+2.59, +3.37]	5.13e-47
Gradient rollout	Foreground energy	-0.12 pp	[-0.30, +0.06]	0.165
Gradient rollout	Pixel AP	-0.68 pp	[-0.94, -0.43]	6.22e-07
Raw CLS	Foreground energy	-0.32 pp	[-0.57, -0.07]	0.00111
Raw CLS	Pixel AP	-0.78 pp	[-1.06, -0.50]	6.34e-10
Attention flow	Deletion AUC (inverted)	+1.35 pp	[+0.69, +2.00]	1.58e-05
Rollout	Deletion AUC (inverted)	+2.96 pp	[+2.33, +3.59]	6.01e-20
Gradient rollout	Deletion AUC (inverted)	+2.81 pp	[+2.15, +3.49]	5.61e-14
Raw CLS	Deletion AUC (inverted)	+2.48 pp	[+1.81, +3.14]	1.55e-12

Table J.6: Paired Direct–Dot insertion, SaCo, and horizontal-flip stability effects for attention-native methods: Attention flow, Rollout, Gradient rollout, Raw CLS, on the same 1,000-image population. Effects are in percentage points with positive favoring Direct; intervals are paired 95% bootstrap intervals and q values use within-family Benjamini–Hochberg correction. Statistical significance alone does not qualify a claim when a required sanity or spatial control fails.

Method	Metric	Direct-dot	95% CI	BH q
Attention flow	Insertion AUC	-1.03 pp	[-1.73, -0.33]	0.00284
Attention flow	SaCo	-0.79 pp	[-1.54, -0.02]	0.0755
Rollout	Insertion AUC	+1.22 pp	[+0.62, +1.81]	8.55e-06
Rollout	SaCo	+0.89 pp	[+0.31, +1.46]	0.00897
Gradient rollout	Insertion AUC	+0.44 pp	[-0.11, +0.99]	0.215
Gradient rollout	SaCo	-0.28 pp	[-0.88, +0.34]	0.572
Raw CLS	Insertion AUC	+0.32 pp	[-0.25, +0.88]	0.698
Raw CLS	SaCo	-0.17 pp	[-0.79, +0.44]	0.676
Attention flow	Flip stability	+1.25 pp	[-1.13, +3.70]	0.246
Rollout	Flip stability	-2.03 pp	[-2.75, -1.29]	1.17e-06
Gradient rollout	Flip stability	+0.09 pp	[-0.30, +0.50]	0.909
Raw CLS	Flip stability	-0.21 pp	[-0.71, +0.30]	0.385

SmoothGrad, and integrated gradients. Integrated gradients also retains all-parameter-randomization correlations of 0.709 for Direct and 0.713 for dot, failing the required parameter-sensitivity control.

J.8 Perturbation Explanations

Perturbation faithfulness does not imply better localization. Patch occlusion favors dot on foreground energy, while the remaining primary localization intervals for TIS and RISE overlap zero. All three methods favor Direct on sign-oriented deletion, but RISE insertion favors dot and patch

Table J.7: Paired Direct–Dot localization, faithfulness, and horizontal-flip stability effects for Transformer-specific methods: Chefer, LeGrad, on the common 1,000-image ImageNet-S population. Effects are in percentage points with positive favoring Direct; intervals are paired 95% bootstrap intervals and q values use within-family Benjamini–Hochberg correction. Statistical significance alone does not qualify a claim when a required sanity or spatial control fails.

Method	Metric	Direct-dot	95% CI	BH q
Chefer	Foreground energy	+0.44 pp	[+0.01, +0.85]	0.0323
Chefer	Pixel AP	+0.27 pp	[-0.14, +0.68]	0.106
LeGrad	Foreground energy	+0.40 pp	[+0.31, +0.48]	1.97e-23
LeGrad	Pixel AP	+0.77 pp	[+0.46, +1.08]	7.54e-09
Chefer	Deletion AUC (inverted)	+2.82 pp	[+2.10, +3.49]	6.28e-13
Chefer	Insertion AUC	+0.32 pp	[-0.26, +0.91]	0.295
Chefer	SaCo	+0.08 pp	[-0.61, +0.73]	0.776
LeGrad	Deletion AUC (inverted)	+2.54 pp	[+1.87, +3.19]	4.57e-14
LeGrad	Insertion AUC	+0.28 pp	[-0.23, +0.78]	0.212
LeGrad	SaCo	+0.21 pp	[-0.40, +0.80]	0.327
Chefer	Flip stability	+3.22 pp	[+1.96, +4.49]	1.03e-06
LeGrad	Flip stability	+0.03 pp	[-0.34, +0.40]	0.236

Table J.8: Paired Direct–Dot localization and sign-oriented deletion effects for class activation mapping methods: Grad-CAM, Grad-CAM++, HiResCAM, XGradCAM, on the common 1,000-image ImageNet-S population. Effects are in percentage points with positive favoring Direct; intervals are paired 95% bootstrap intervals and q values use within-family Benjamini–Hochberg correction. Statistical significance alone does not qualify a claim when a required sanity or spatial control fails.

Method	Metric	Direct-dot	95% CI	BH q
Grad-CAM	Foreground energy	-1.05 pp	[-1.53, -0.58]	6.21e-05
Grad-CAM	Pixel AP	-5.35 pp	[-6.15, -4.57]	3.06e-43
Grad-CAM++	Foreground energy	-2.63 pp	[-4.37, -0.89]	0.00071
Grad-CAM++	Pixel AP	-2.91 pp	[-4.56, -1.26]	0.000447
HiResCAM	Foreground energy	-25.27 pp	[-26.58, -23.95]	2.9e-136
HiResCAM	Pixel AP	-24.67 pp	[-26.23, -23.19]	4.08e-130
XGradCAM	Foreground energy	-1.68 pp	[-3.32, +0.01]	0.055
XGradCAM	Pixel AP	-1.68 pp	[-3.38, +0.05]	0.0579
Grad-CAM	Deletion AUC (inverted)	+1.10 pp	[+0.39, +1.82]	0.00284
Grad-CAM++	Deletion AUC (inverted)	-0.76 pp	[-1.91, +0.36]	0.106
HiResCAM	Deletion AUC (inverted)	-15.80 pp	[-17.10, -14.50]	1.29e-94
XGradCAM	Deletion AUC (inverted)	+0.20 pp	[-1.03, +1.42]	0.688

occlusion has a small dot-favored SaCo effect. With only one of four localization controls passing for patch occlusion and TIS and none for RISE, these perturbation results cannot establish a localization or shared-mechanism claim.

Table J.9: Paired Direct–Dot insertion, SaCo, and horizontal-flip stability effects for class activation mapping methods: Grad-CAM, Grad-CAM++, HiResCAM, XGradCAM, on the same 1,000-image population. Effects are in percentage points with positive favoring Direct; intervals are paired 95% bootstrap intervals and q values use within-family Benjamini–Hochberg correction. Statistical significance alone does not qualify a claim when a required sanity or spatial control fails.

Method	Metric	Direct-dot	95% CI	BH q
Grad-CAM	Insertion AUC	-0.41 pp	[-0.95, +0.15]	0.104
Grad-CAM	SaCo	+0.01 pp	[-0.71, +0.72]	0.801
Grad-CAM++	Insertion AUC	-3.04 pp	[-4.19, -1.91]	2.35e-08
Grad-CAM++	SaCo	-0.79 pp	[-2.07, +0.46]	0.405
HiResCAM	Insertion AUC	-18.30 pp	[-19.67, -16.93]	1.86e-111
HiResCAM	SaCo	-8.75 pp	[-9.93, -7.55]	1.1e-43
XGradCAM	Insertion AUC	-2.27 pp	[-3.47, -1.12]	0.000381
XGradCAM	SaCo	-0.20 pp	[-1.45, +1.03]	0.509
Grad-CAM	Flip stability	-2.67 pp	[-3.54, -1.83]	3e-09
Grad-CAM++	Flip stability	-2.70 pp	[-8.85, +3.54]	0.313
HiResCAM	Flip stability	-1.31 pp	[-3.05, +0.40]	0.213
XGradCAM	Flip stability	+1.68 pp	[-4.99, +8.49]	0.571

Table J.10: Paired Direct–Dot localization and sign-oriented deletion effects for input-gradient methods: Input×Gradient, Integrated gradients, Saliency, SmoothGrad, on the common 1,000-image ImageNet-S population. Effects are in percentage points with positive favoring Direct; intervals are paired 95% bootstrap intervals and q values use within-family Benjamini–Hochberg correction. Statistical significance alone does not qualify a claim when a required sanity or spatial control fails.

Method	Metric	Direct-dot	95% CI	BH q
Input×Gradient	Foreground energy	+0.10 pp	[-0.37, +0.56]	0.349
Input×Gradient	Pixel AP	-0.00 pp	[-0.58, +0.57]	0.348
Integrated gradients	Foreground energy	-0.09 pp	[-0.26, +0.08]	0.256
Integrated gradients	Pixel AP	-0.38 pp	[-0.72, -0.04]	0.0759
Saliency	Foreground energy	-0.34 pp	[-0.84, +0.16]	0.482
Saliency	Pixel AP	-0.66 pp	[-1.23, -0.10]	0.309
SmoothGrad	Foreground energy	-0.07 pp	[-0.35, +0.21]	0.703
SmoothGrad	Pixel AP	+0.00 pp	[-0.48, +0.48]	0.889
Input×Gradient	Deletion AUC (inverted)	+1.73 pp	[+1.08, +2.38]	5.22e-08
Integrated gradients	Deletion AUC (inverted)	+1.95 pp	[+1.34, +2.59]	1.61e-07
Saliency	Deletion AUC (inverted)	+1.67 pp	[+1.05, +2.31]	2.38e-06
SmoothGrad	Deletion AUC (inverted)	+1.82 pp	[+1.18, +2.45]	5.22e-07

J.9 Sanity and Stability

Randomization failures veto otherwise favorable effects. Integrated gradients remains strongly correlated with its trained map after all parameters are randomized, with rank correlations of 0.709 for Direct and 0.713 for dot, and is therefore marked as failing the sanity audit. Stability also varies sharply across methods, with especially low transformation consistency for XGradCAM and RISE. Neither visual plausibility nor stability can substitute for parameter sensitivity and faithfulness.

Table J.11: Paired Direct–Dot insertion, SaCo, and horizontal-flip stability effects for input-gradient methods: Input×Gradient, Integrated gradients, Saliency, SmoothGrad, on the same 1,000-image population. Effects are in percentage points with positive favoring Direct; intervals are paired 95% bootstrap intervals and q values use within-family Benjamini–Hochberg correction. Statistical significance alone does not qualify a claim when a required sanity or spatial control fails.

Method	Metric	Direct-dot	95% CI	BH q
Input×Gradient	Insertion AUC	+0.16 pp	[-0.42, +0.75]	0.432
Input×Gradient	SaCo	+0.35 pp	[-0.24, +0.96]	0.632
Integrated gradients	Insertion AUC	-0.18 pp	[-0.74, +0.39]	0.563
Integrated gradients	SaCo	-0.18 pp	[-0.88, +0.52]	0.841
Saliency	Insertion AUC	-0.41 pp	[-1.01, +0.18]	0.22
Saliency	SaCo	-0.12 pp	[-0.73, +0.48]	0.417
SmoothGrad	Insertion AUC	-0.56 pp	[-1.13, +0.01]	0.0541
SmoothGrad	SaCo	-0.09 pp	[-0.71, +0.56]	0.4
Input×Gradient	Flip stability	+0.37 pp	[-0.59, +1.34]	0.909
Integrated gradients	Flip stability	+0.04 pp	[-0.32, +0.39]	0.291
Saliency	Flip stability	+0.37 pp	[-0.48, +1.24]	0.454
SmoothGrad	Flip stability	+0.97 pp	[+0.39, +1.58]	0.0116

Table J.12: Paired Direct–Dot localization, faithfulness, and horizontal-flip stability effects for perturbation methods: Patch occlusion, RISE, TIS, on the common 1,000-image ImageNet-S population. Effects are in percentage points with positive favoring Direct; intervals are paired 95% bootstrap intervals and q values use within-family Benjamini–Hochberg correction. Statistical significance alone does not qualify a claim when a required sanity or spatial control fails.

Method	Metric	Direct-dot	95% CI	BH q
Patch occlusion	Foreground energy	-0.41 pp	[-0.66, -0.15]	0.00666
Patch occlusion	Pixel AP	-0.23 pp	[-0.87, +0.41]	0.685
RISE	Foreground energy	-0.17 pp	[-0.21, -0.12]	2.7e-13
RISE	Pixel AP	-0.97 pp	[-1.24, -0.70]	1.98e-12
TIS	Foreground energy	+0.26 pp	[+0.03, +0.48]	0.00431
TIS	Pixel AP	+0.07 pp	[-0.30, +0.43]	0.77
Patch occlusion	Deletion AUC (inverted)	+2.61 pp	[+1.79, +3.46]	2.79e-10
Patch occlusion	Insertion AUC	+0.54 pp	[-0.10, +1.18]	0.286
Patch occlusion	SaCo	-0.01 pp	[-0.01, +0.00]	0.0376
RISE	Deletion AUC (inverted)	+0.47 pp	[-0.23, +1.15]	0.199
RISE	Insertion AUC	-0.54 pp	[-1.09, +0.01]	0.0376
RISE	SaCo	+0.32 pp	[-0.01, +0.64]	0.128
TIS	Deletion AUC (inverted)	+1.58 pp	[+0.89, +2.25]	1.51e-06
TIS	Insertion AUC	+0.49 pp	[-0.04, +1.02]	0.107
TIS	SaCo	+1.66 pp	[+1.18, +2.16]	1.08e-09
Patch occlusion	Flip stability	+3.42 pp	[+2.07, +4.81]	7.29e-07
RISE	Flip stability	-0.58 pp	[-1.22, +0.07]	0.149
TIS	Flip stability	+1.77 pp	[+0.80, +2.75]	0.00312

Table J.13: Method-level sanity and claim-control summary for all 17 explainers. Sanity passes only when both checkpoints respond to cascading parameter randomization; Controls reports the number of localization metrics, out of four, that beat every required spatial baseline. Full-set direction applies to the six methods evaluated on all 4,580 masks, and Stronger claims counts effects admitted by the complete qualification rule. Any required failure vetoes positive wording.

Method	Sanity	Controls	Full-set direction	Stronger claims
Raw CLS	pass	0/4	fail	3
Rollout	pass	0/4	fail	7
Attention flow	pass	0/4	n/a	0
Gradient rollout	pass	1/4	fail	3
Chefer	pass	3/4	n/a	6
LeGrad	pass	3/4	fail	4
Grad-CAM	pass	2/4	fail	2
Grad-CAM++	pass	0/4	n/a	0
HiResCAM	pass	0/4	n/a	0
XGradCAM	pass	0/4	n/a	0
Saliency	pass	0/4	n/a	2
Input×Gradient	pass	0/4	fail	2
SmoothGrad	pass	0/4	n/a	4
Integrated gradients	fail	0/4	n/a	0
Patch occlusion	pass	1/4	n/a	4
TIS	pass	1/4	n/a	4
RISE	pass	0/4	n/a	0

J.10 Balanced Outcomes and Failure Cases

Qualitative heterogeneity. The shared-normalization gallery exposes Direct-favored, dot-favored, near-equal, and low-localization cases rather than presenting only favorable examples. Because these categories are selected using an aggregate of rollout, Grad-CAM, and TIS foreground energy, they illustrate the observed heterogeneity rather than a population effect. Quantitative conclusions follow the common-population metrics and controls above.

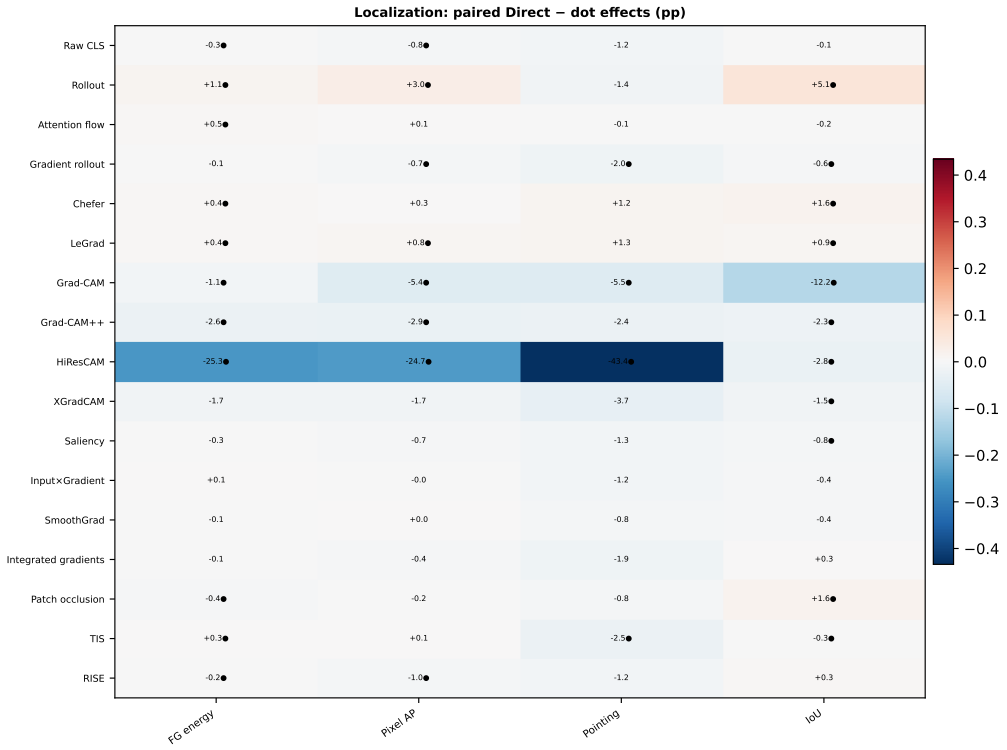


Figure J.3: Localization effects for all 17 explainers on the common 1,000-image ImageNet-S population. Columns report foreground energy, pixel average precision, pointing-game accuracy, and mean intersection over union; cells are paired Direct-minus-Dot effects in percentage points with positive favoring Direct. Statistical markers encode paired 95% intervals and within-family Benjamini–Hochberg decisions, and full-4,580-mask direction is required for the six low-cost methods. Center-prior failures and dot-favored CAM results remain visible rather than being filtered from the comparison.

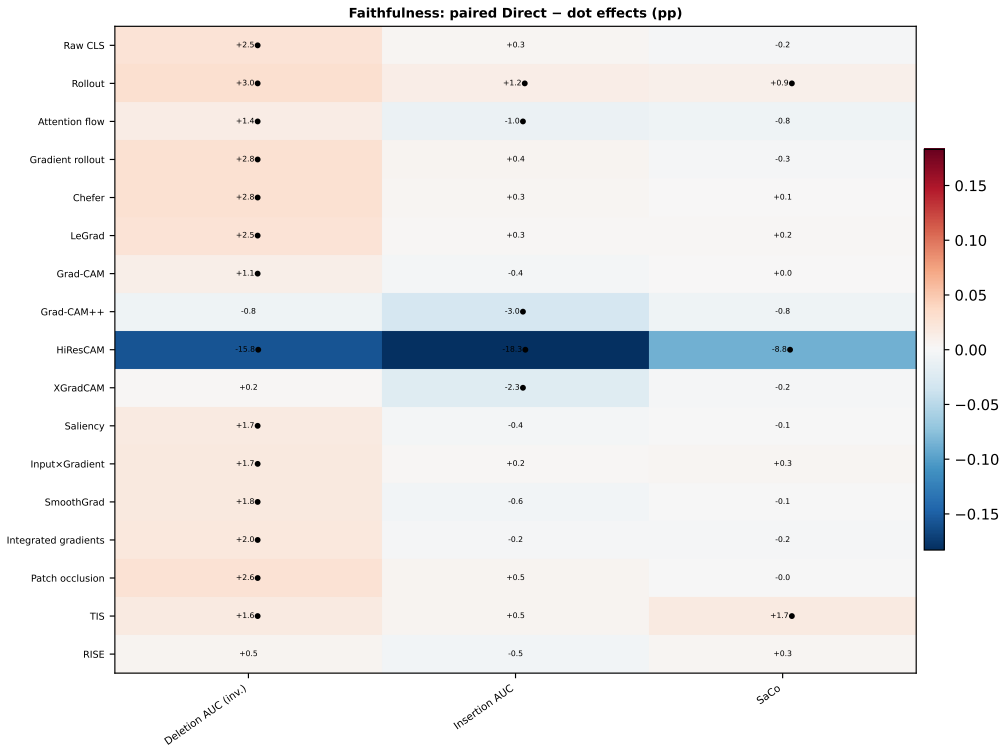


Figure J.4: Faithfulness effects for every explainer on the common 1,000-image population. The panels report insertion area under the curve, sign-oriented deletion area under the curve, and the Salience-guided Faithfulness Coefficient (Wu et al., 2024); all paired Direct-minus-Dot effects are oriented so positive favors Direct. Markers summarize paired 95% bootstrap intervals and within-family Benjamini–Hochberg decisions. The metric families remain separate because favorable deletion does not imply favorable insertion, SaCo, localization, or stability.

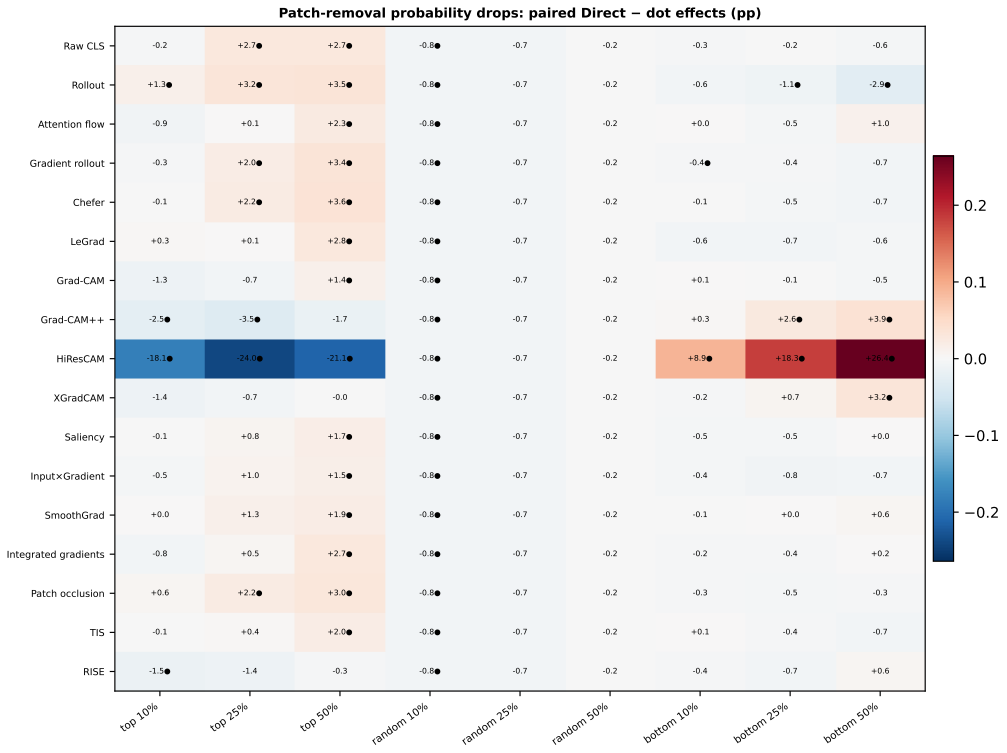


Figure J.5: Direct-minus-Dot effects for model-specific patch removal on 1,000 images. Each checkpoint supplies its own top, random, and bottom patch ranking, evaluated at 10%, 25%, and 50% removal; cells report differences in original-class probability drop, with positive indicating larger damage for Direct. Paired intervals and corrected decisions are shown for every method and budget. Because the removed regions differ between checkpoints, these are faithfulness comparisons rather than shared spatial interventions.

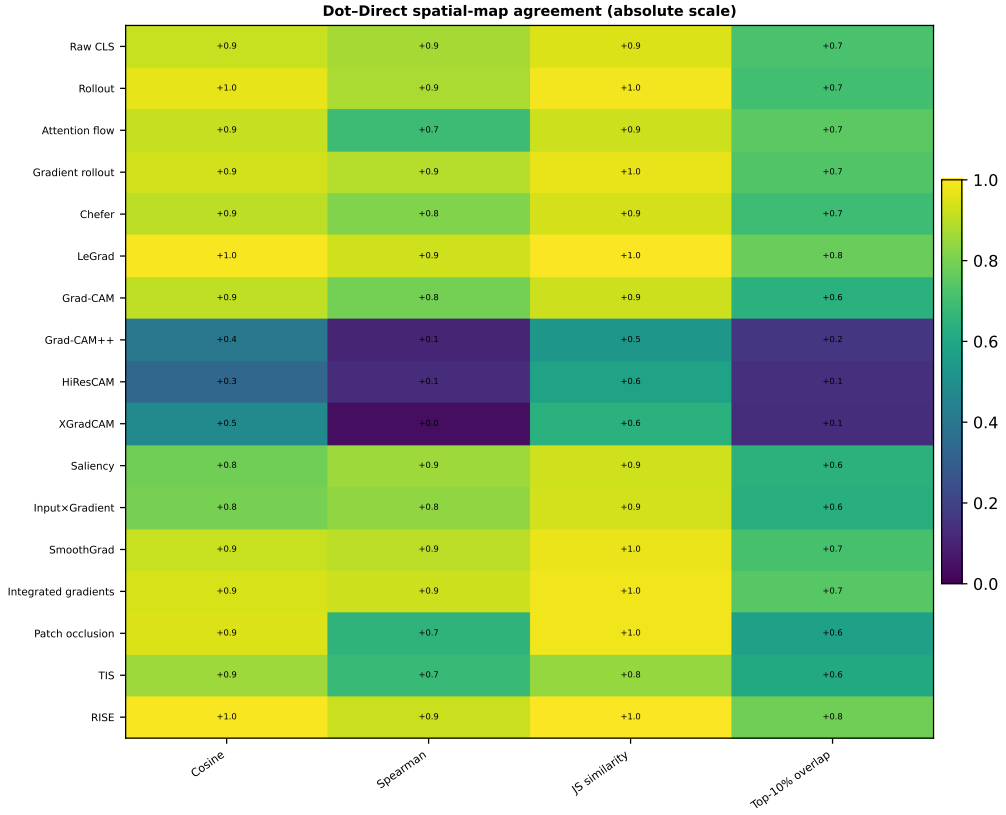


Figure J.6: Absolute Dot–Direct explanation-map agreement for all 17 methods on the common 1,000-image population. Columns report map cosine, Spearman rank correlation, Jensen–Shannon similarity, and top-10% patch overlap, with higher values indicating more similar spatial rankings or mass. These are symmetric agreement measures rather than Direct-minus-Dot advantage effects, and no significance marker converts agreement into faithfulness. High or low agreement therefore cannot determine which checkpoint provides the better explanation.

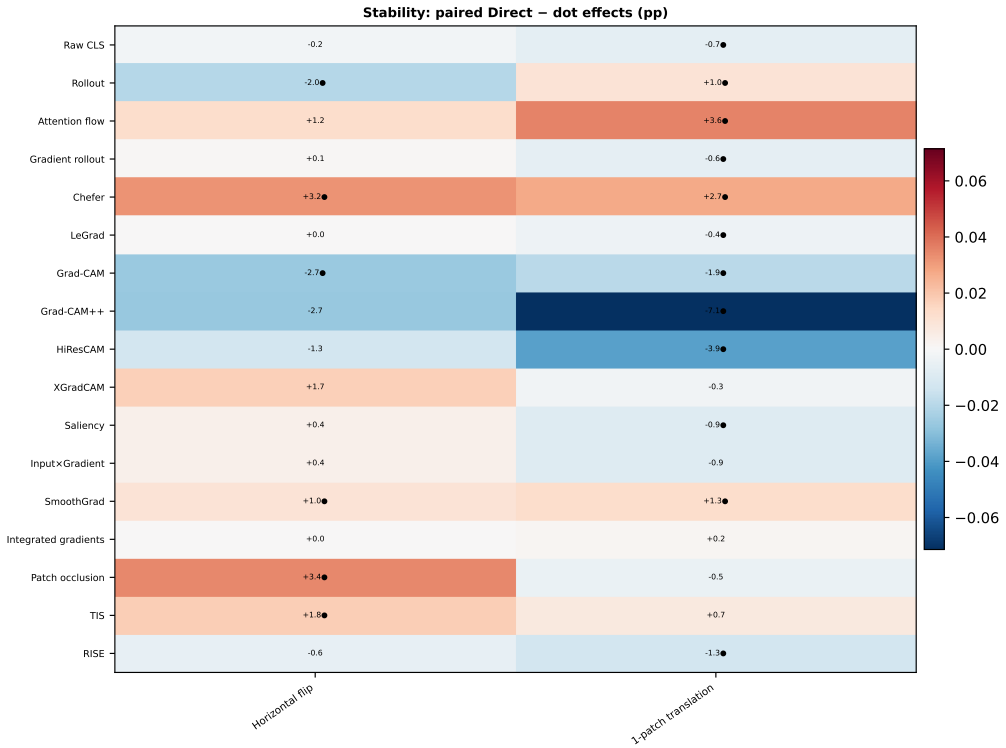


Figure J.7: Transformation-stability comparison on the fixed 250-image subset. For each explainer, horizontal flips and one-patch translations are inverted back to the original coordinates before rank consistency is measured. Cells show paired Direct-minus-Dot stability effects, with positive favoring Direct, and markers encode corrected paired decisions. Stability describes response to these two transformations and is neither a localization nor a faithfulness guarantee.

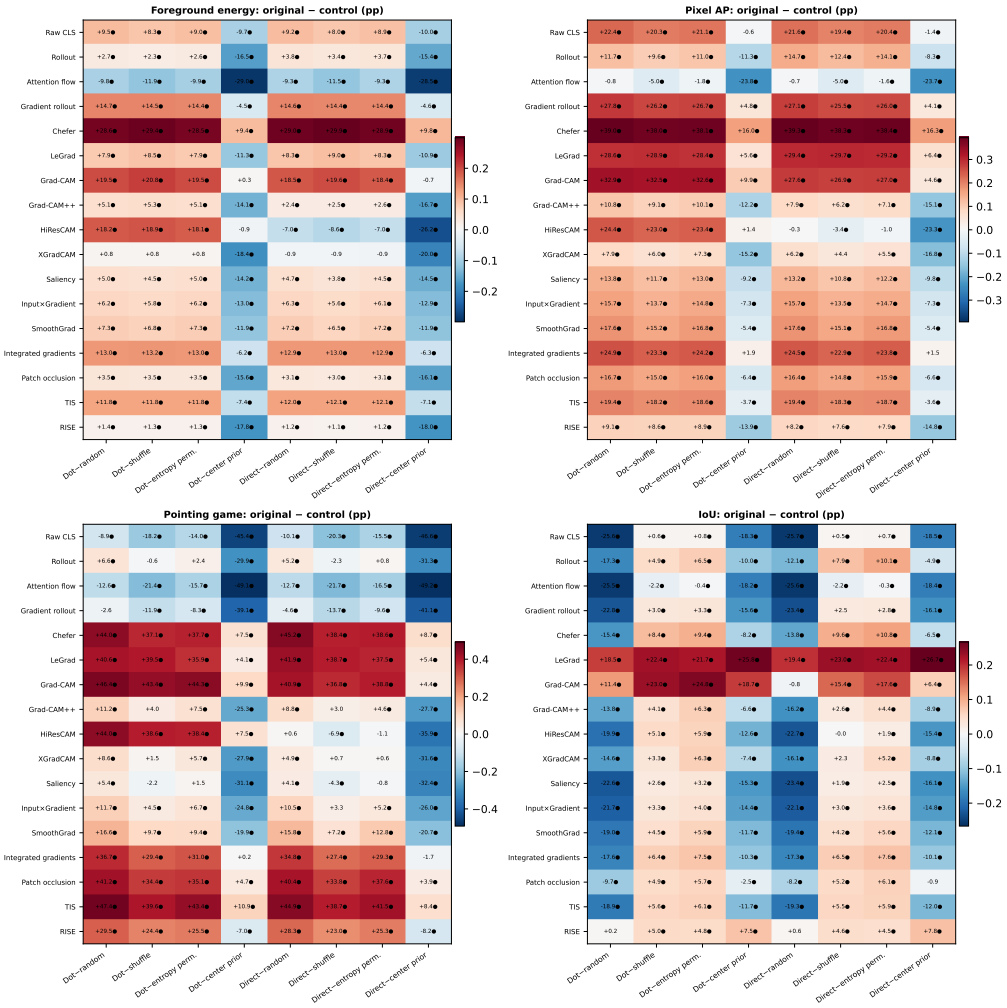


Figure J.8: Spatial localization controls for both checkpoints on all 4,580 valid ImageNet-S masks. Each matrix compares the original map with random maps, spatial shuffle, entropy-preserving permutation, and the center prior under foreground energy, pixel average precision, pointing game, and intersection over union; positive effects favor the original map. Filled circles mark favorable paired intervals with corrected significance, and every failure remains visible. A localization result is control-qualified only when the original map beats every required baseline, including the center prior.

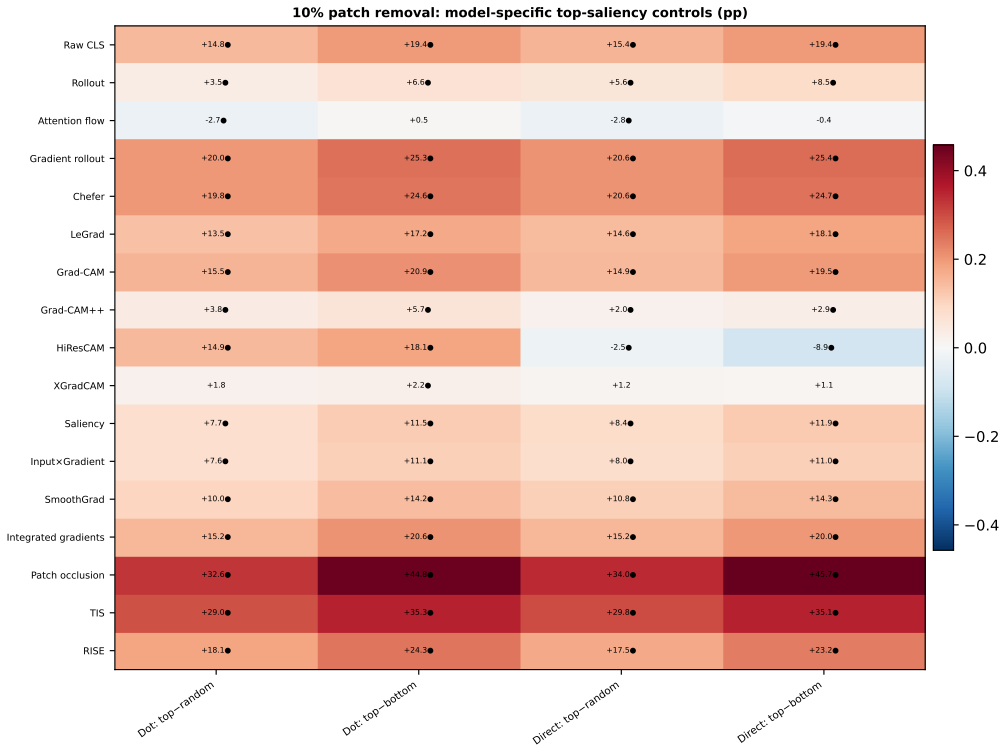


Figure J.9: Faithfulness controls at the 10% patch-removal budget for every checkpoint–explainer pair. Cells compare original-class probability drop for top-ranked patches with random- and bottom-ranked patches using each model's own map; positive top-minus-control effects indicate that the ranking identifies more consequential regions. Markers encode paired 95% intervals and within-family corrected decisions. Passing is specific to the named checkpoint, explainer, and removal budget.

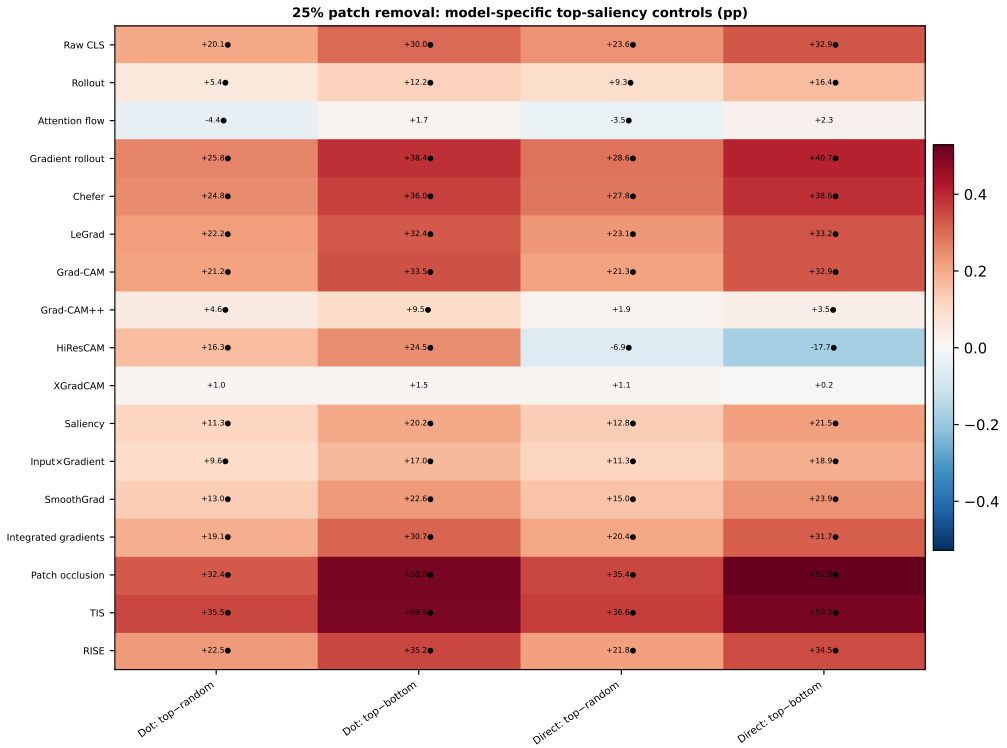


Figure J.10: Faithfulness controls at the 25% patch-removal budget for every checkpoint–explainer pair. Cells compare own-model top-ranked removal with random and bottom removal in original-class probability drop, so positive effects favor the explainer's ranking. Paired 95% intervals and within-family corrected markers retain both passes and failures. The test supports only the named checkpoint, method, and budget, not shared salient regions.

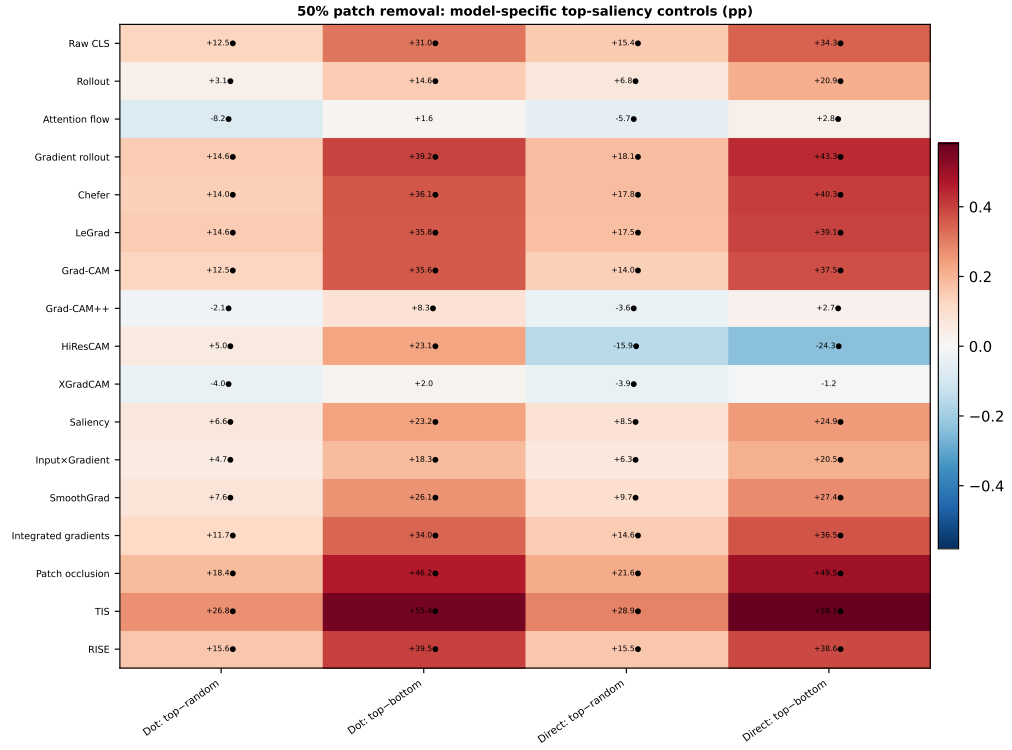


Figure J.11: Faithfulness controls at the 50% patch-removal budget for every checkpoint–explainer pair. Top-minus-random and top-minus-bottom effects use each checkpoint's own ranking and measure additional original-class probability drop, with positive values indicating greater predictive consequence. Markers summarize paired 95% intervals and corrected decisions; failed rows remain in the matrix. Results at this largest budget cannot be generalized to lower removal fractions or cross-model spatial agreement.

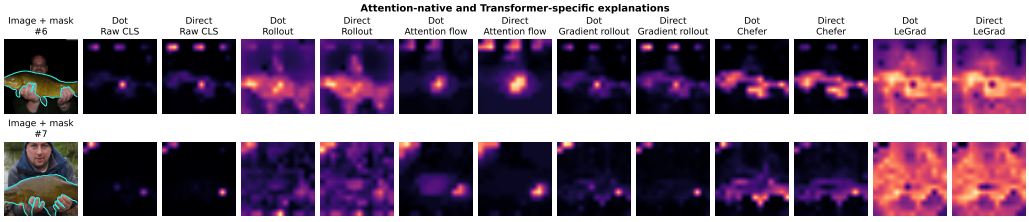


Figure J.12: Attention-native and Transformer-specific explanations for raw CLS attention, attention rollout, attention flow, gradient-weighted rollout, Chefer relevance, and LeGrad (Abnar & Zuidema, 2020; Chefer et al., 2021; Bousselham et al., 2025). Each row uses the same ground-truth target for the paired Dot and Direct checkpoints, the input carries the ImageNet-S mask contour, and each method pair shares its normalization; examples follow the fixed gallery selection rather than favorable outcomes. Chefer uses its own residual relevance recurrence rather than a gradient-rollout alias. The maps are illustrative, and rollout's center-prior failure plus the mixed insertion effects remain controlling evidence.

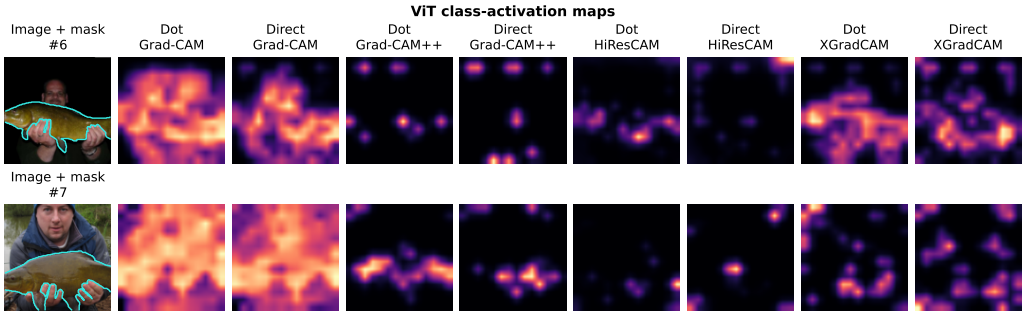


Figure J.13: Class activation maps from Grad-CAM, Grad-CAM++, HiResCAM, and XGradCAM (Selvaraju et al., 2017; Chattopadhyay et al., 2018; Draelos & Carin, 2020; Fu et al., 2020) using the final Transformer block's pre-attention normalization, removal of the CLS token, and the same 14-by-14 reshape for both checkpoints. Dot and Direct use the same target class, shared pairwise normalization, fixed example selection, and ImageNet-S contours. The gallery retains four constant Grad-CAM++ maps and the substantial dot-favored localization effects of the CAM family. These visualizations therefore document mixed outcomes rather than a general Direct advantage.

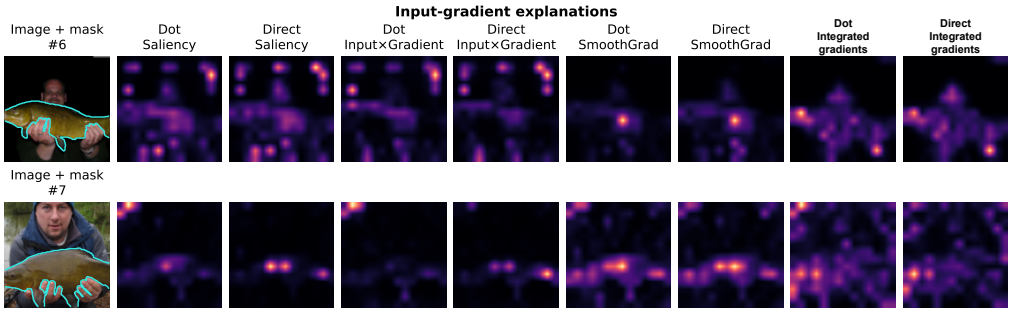


Figure J.14: Input-gradient explanations from saliency, Input×Gradient, SmoothGrad, and integrated gradients (Smilkov et al., 2017; Sundararajan et al., 2017) for the same target class and fixed examples. Each Dot/Direct pair shares normalization and the input includes the ImageNet-S contour; positive class evidence is used for localization while signed behavior is summarized separately. SmoothGrad includes dot-favored localization effects, and integrated gradients fails the parameter-randomization sanity criterion. Consequently, visual plausibility in this gallery cannot qualify an otherwise failed method.

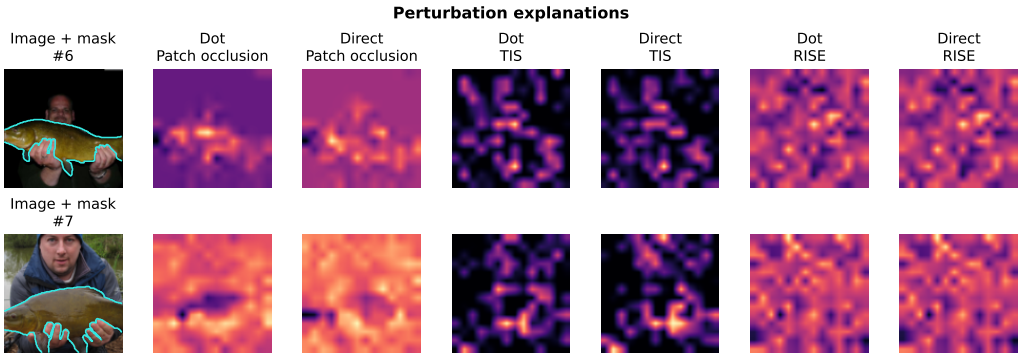


Figure J.15: Perturbation explanations from patch occlusion, Transformer Input Sampling (TIS), and token-level RISE (Englebert et al., 2023; Petsiuk et al., 2018) for the shared target class and fixed gallery examples. Occlusion removes one 16-by-16 patch against a blurred baseline, whereas TIS and RISE use 512 deterministic token masks with keep probability 0.5; paired maps share normalization and mask contours. TIS and RISE perturb patch embeddings rather than zeroing image pixels. Their uneven localization controls and the dot-favored RISE insertion effect prevent these maps from supporting a localization-mechanism claim.

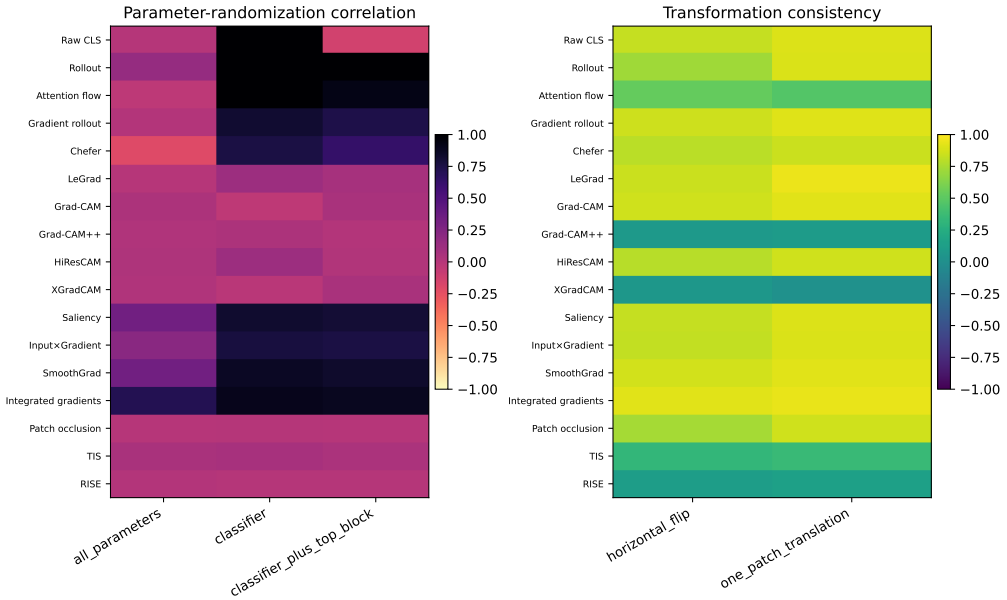


Figure J.16: Sanity and stability audit for the complete explainer suite. Cascading parameter randomization uses a fixed 100-image subset, while horizontal-flip and one-patch-translation consistency use 250 images. Filled markers denote retained sensitivity or corrected stability effects and failure symbols identify a required sanity violation. Integrated gradients remains highly correlated after full randomization and therefore fails, while XGradCAM and RISE show especially weak transformation consistency. Neither stability nor visual plausibility substitutes for parameter sensitivity and faithfulness.

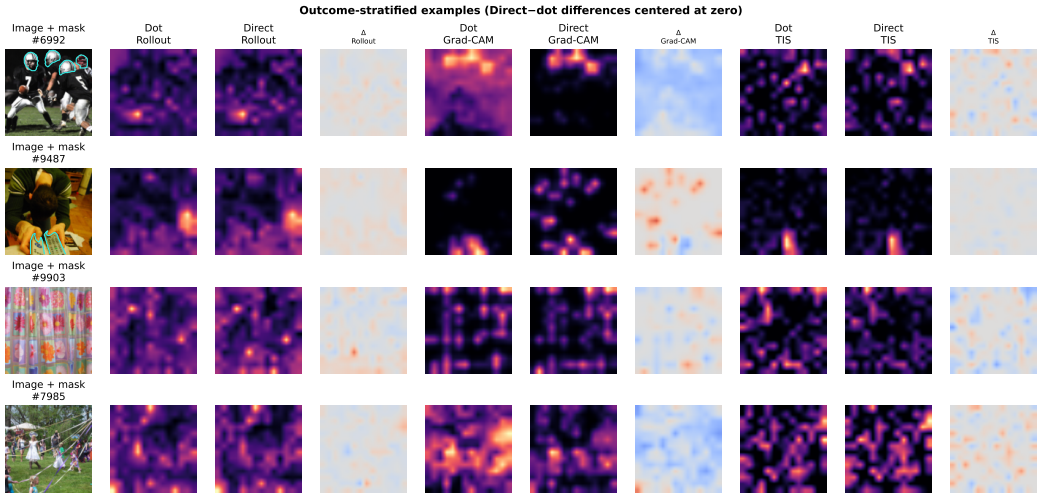


Figure J.17: Balanced qualitative outcomes selected from the common population by an aggregate of rollout, Grad-CAM, and Transformer Input Sampling foreground energy. The rows intentionally include Direct-favored, dot-favored, near-equal, and typical low-localization cases; inputs show ImageNet-S contours and paired maps use the same target class and shared normalization. The diverging difference map is Direct minus Dot with zero centered. Because the categories depend on observed outcomes, the gallery illustrates heterogeneity and selection risk rather than estimating a population effect.

K HISTORICAL STAGE 1 AND STAGE 2 EVIDENCE

Historical scope. Stage 1 and Stage 2 are earlier diagnostic audits, distinct from the training phases of the primary ImageNet conversion. Stage 1's post-hoc gate failed because only the functional family passed. Its comparisons include the independent Tiny dot/Direct pair, identity swap, wrong-sign and failed-operator controls, and two archived ViT-Small pairs. The ViT-Small/CIFAR pair has an incorrect 1,000-output classifier and is excluded from the main evidence. Stage 2 localization is superseded by the controlled Stage 3 population.

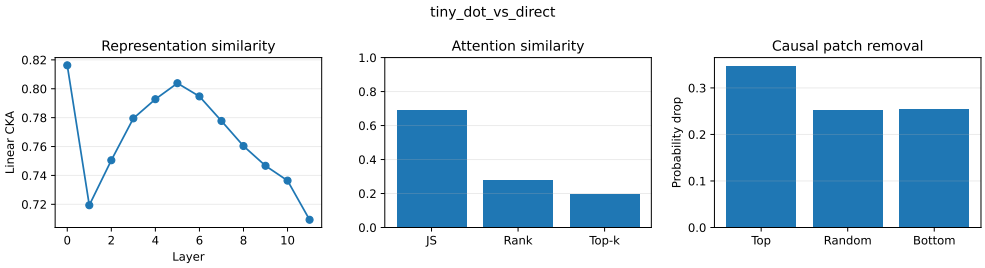


Figure K.1: Historical Tiny Dot-versus-Direct comparison on CIFAR-100, including attention similarity, prediction/logit behavior, representation comparisons, and the original causal controls. The frozen checkpoints come from the same single-seed exploratory setting as the trajectory plot. Only the functional family passed the post-hoc "two-of-three" Stage 1 gate, while the causal family failed. The figure is retained to disclose that negative result and is not used as authoritative Stage 3 mechanism evidence.

K.1 SUPERSEDED BIAS-DELETED KD CHECKPOINT EVIDENCE

Superseded checkpoint. The first ImageNet-1K Direct checkpoint used label-aware knowledge distillation with fused-QKV-bias deletion (runs/final_dk_small_imagenet/checkpoints/best_kd.ckpt; evidence bundle inspect/outputs/stage3). On all 50000 validation images it measured 74.634% for the Dot checkpoint and 74.380% for the Direct checkpoint (-0.254 pp, 95% CI $[-0.486, -0.014]$), it omitted 12 fused-QKV bias tensors, and restoring Q, K, V, QK, or all QKV biases changed accuracy by -2.8, -2.5, -12.4, -7.7, and -21.0 pp respectively—a structural confound that prevented a clean QK-only claim. The bias-preserving legacy checkpoint in Appendix D superseded it for the historical mechanism audits. Neither is the primary conversion endpoint, and the bias-deleted results are not pooled with either bias-preserving checkpoint's evidence.

Table K.1: Historical Stage 1 and Stage 2 evidence retained for disclosure. The table reports the single-seed Tiny comparisons, wrong-sign and failed-operator controls, the excluded incorrect-head checkpoint, and the superseded localization study. Stage 1 failed its post-hoc gate, so no row is authoritative mechanism evidence.

Run	Attention Spearman	Prediction agreement	Logit / representation	Causal / localization	Status
Stage 1 post-hoc gate	–	–	functional family only	causal family failed	failed; historical non-authoritative
tiny_dot_vs_direct	0.279124	0.647887	cosine 0.717; mean CKA 0.766	candidate top/random drop 0.348/0.252	historical non-authoritative
tiny_identity_swap	0.545957	0.349721	cosine 0.576; mean CKA 0.695	candidate top/random drop 0.396/0.270	historical non-authoritative
tiny_wrong_sign	0.0514864	0.0333773	cosine 0.108; mean CKA 0.214	candidate top/random drop 0.010/0.003	historical non-authoritative
tiny_failed_abs	0.071066	0.0705692	cosine 0.144; mean CKA 0.294	candidate top/random drop 0.018/0.011	historical non-authoritative
small_cifar_dot_vs_direct	0.819977	0.867536	cosine 0.997; mean CKA 0.972	candidate top/random drop 0.458/0.352	wrong classifier head; excluded
small_imagenet_dot_vs_direct	0.768481	0.876499	cosine 0.960; mean CKA 0.900	candidate top/random drop 0.228/0.093	historical non-authoritative
Stage2 imagenet_s_dot_vs_direct	0.769013	0.893897	cosine 0.961; mean CKA 0.904	pointing dot/Direct 0.421/0.596	superseded by controlled Stage 3

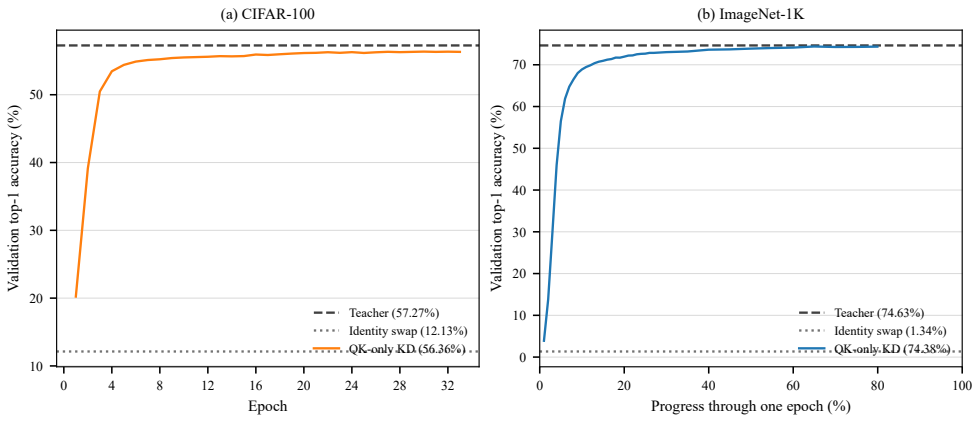


Figure K.2: Historical knowledge-distillation adaptation curves for the superseded bias-deleted Direct checkpoint (ImageNet-1K, validation top-1 accuracy versus progress through one epoch). These runs trained the Q/K rows with label-aware distillation and deleted the fused-QKV biases; the results are superseded for the legacy audit by the bias-preserving checkpoint in Appendix D. They are not training curves for the primary conversion study.

L TRAINING-SEMANTICS DIAGNOSTIC

Diagnostic protocol. The 500-step diagnostic starts Direct, raw minimum, and a tie-controlled gauge-fixed intersection branch from one ViT-Tiny/CIFAR-100 initialization and feeds identical minibatches in bf16. It records steps 0, 1, 10, 50, 100, 250, and 500. This is an implementation diagnostic with no accuracy acceptance target, not a replication experiment.

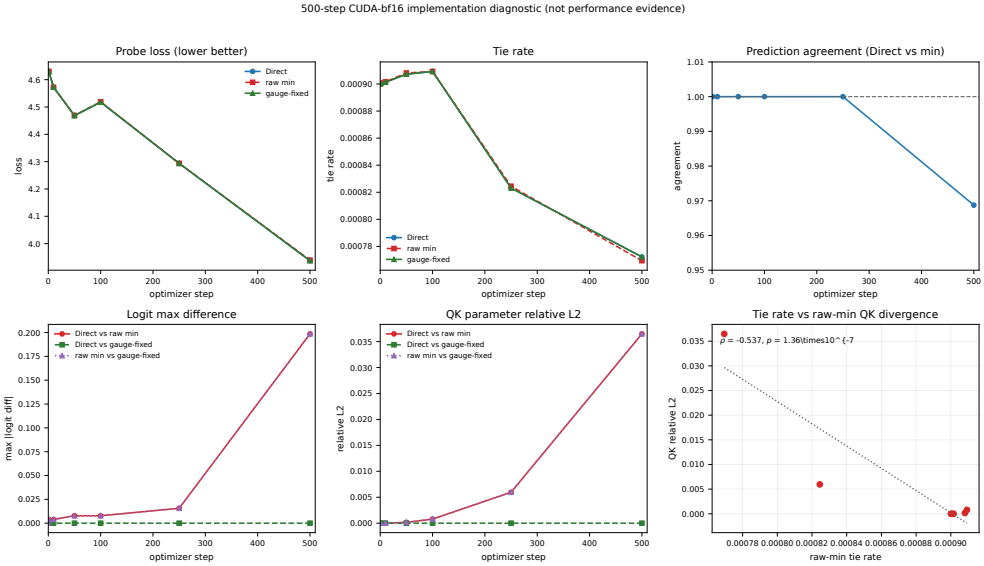


Figure L.1: Single-run Direct/minimum training-semantics diagnostic from one ViT-Tiny/CIFAR-100 initialization: Direct, raw minimum, and tie-controlled gauge-fixed intersection receive identical minibatches for 500 bf16 steps, with losses, tie rates, logits, and Q/K parameter divergence recorded at seven checkpoints (linear step axis). Gauge-fixed intersection tracks Direct throughout; raw-minimum divergence is negatively associated with tie frequency ($\rho = -0.537$, $p = 1.36 \times 10^{-7}$; panel (f)), and Direct-vs-min prediction agreement stays at 1 until it drops to 0.96875 at step 500. This is a numerical training-semantics diagnostic without a performance acceptance target, so it does not resolve the historical 7.33% trajectory or establish equal training dynamics.

Training dynamics remain causally unresolved. The gauge-fixed intersection branch matches Direct at every recorded step, whereas the raw-minimum branch accumulates a QK parameter relative difference of 0.036491 and falls to 96.875% prediction agreement by step 500. Raw-min divergence is negatively, rather than positively, associated with tie frequency ($\rho = -0.537$, $p = 1.36 \times 10^{-7}$), so this diagnostic does not support ties as the cause of the historical min failure. Because this is one 500-step bf16 trace from one initialization with no accuracy acceptance target, the historical 7.33% trajectory remains causally unresolved. The experiment therefore establishes neither identical training dynamics nor intrinsic inferiority of the raw-min operator.

Table L.1: Seven recorded points from the single 500-step Direct/minimum training-semantics diagnostic. Loss, prediction agreement, attention/logit drift, Q/K divergence, and tie behavior compare Direct, raw minimum, and gauge-fixed intersection under identical minibatches. The trace has no accuracy acceptance target and leaves the historical minimum anomaly causally unresolved.

Step	Loss: Direct / raw min / gauge-fixed	Tie rate: Direct / raw min	Direct-vs-min agreement	Logit max difference	QK parameter relative L2	Gauge-fixed control
0	4.6287 / 4.6287 / 4.6287	9.001e-04 / 9.001e-04	1	0	0	agreement 1.000; logit diff 0.0e+00
1	4.6302 / 4.6302 / 4.6302	9.006e-04 / 9.006e-04	1	0.00390625	0	agreement 1.000; logit diff 0.0e+00
10	4.5723 / 4.5723 / 4.5723	9.013e-04 / 9.017e-04	1	0.00390625	4.13750e-06	agreement 1.000; logit diff 0.0e+00
50	4.4684 / 4.4693 / 4.4684	9.072e-04 / 9.081e-04	1	0.0078125	0.000154893	agreement 1.000; logit diff 0.0e+00
100	4.5178 / 4.5191 / 4.5178	9.093e-04 / 9.093e-04	1	0.0078125	0.000788869	agreement 1.000; logit diff 0.0e+00
250	4.2936 / 4.2936 / 4.2936	8.231e-04 / 8.245e-04	1	0.015625	0.00595721	agreement 1.000; logit diff 0.0e+00
500	3.9373 / 3.9396 / 3.9373	7.722e-04 / 7.694e-04	0.96875	0.198624	0.036491	agreement 1.000; logit diff 0.0e+00

M VISION-LANGUAGE TRANSFER EXPERIMENT REPORT

Models and data. The transfer is repeated on MiniMind-V (SigLIP2 vision encoder plus a decoder-only LLM backbone; MoE 200M-A65M and dense 65M variants) and the matching plain MiniMind LLM (65M, no vision branch). Training data is the SFT conversation parquet of the MiniMind release: a 200,000-row training pool, a 2,000-row selection split, and a 2,000-row evaluation split; the plain-LLM runs use the text-only subset of the same parquet (228,544 rows) with the same 200k/2k/2k split. Every reported selection or evaluation CE covers the full corresponding 2,000-row split; training-curve CE values are minibatch losses. A legacy evaluator field labeled "rows" counted batches (500 four-example VLM batches or 250 eight-example LLM batches), not examples. All runs are single-seed development evidence with the same evidential status as the ImageNet development run; no confirmatory statistics are attached to them.

Replacement procedure. We reuse the operator-replacement and Q/K-only optimization template, with the transfer-specific objective and schedule below. Direct scoring replaces the dot score in the chosen modules of the pretrained checkpoint (weight-identical swap, forward gate error $< 10^{-7}$). Only Q/K weights in the selected modules are optimized; every other tensor, including non-Q/K vision-encoder tensors, remains frozen. The KD-mixture objective uses attention soft-CE at temperature 1 plus output distillation at $T = 3$ with T^2 scaling, $\alpha = 0.9$, and dense mix 0.5. The ordinary-arm schedule uses learning rate 10^{-4} with cosine decay and 100 warm-up steps, gradient clipping at 1.0, bf16 autocast, micro-batch 4 $\times$ gradient accumulation 2, sequence length 384, selection every 1,000 optimizer steps, patience five with minimum delta 5×10^{-3}, and a 20,000-step cap. The MoE arm is an unmatched development-time exception: after its initial 20,000 steps, it resumed once from the retained checkpoint at learning rate 2×10^{-5} with a 40,000 global-step ceiling; it stopped at step 39,000 after five non-improving checks and retained step 34000. Trainable tensor counts are 16 (decoder-only), 24 (SigLIP encoder only), and 40 (whole model).

Complete run table. Table M.1 reports every principal transfer run and the E1 dummy-image training control: selection CE and selected step, evaluation CE, the same-split evaluation delta over the corresponding dot baseline, and the integrity result (exactly the arm's Q/K tensors change; final checkpoint equals the selected one; all other tensors bit-identical).

Table M.1: Complete MiniMind transfer record (single seeds, development evidence). "Q/K" is the number of changed Q/K projection tensors (integrity). Baseline CE and Eval CE are measured on the corresponding held-out 2,000-row evaluation split; $\Delta\mathrm{CE} = \mathrm{Eval\ CE} - \mathrm{Baseline\ CE}$. Endpoint CEs and deltas are independently rounded to three decimals; deltas are computed from the unrounded endpoint values, so subtracting displayed CEs can differ by 0.001. Sel. CE is measured on the disjoint 2,000-row selection split.

Run	Baseline eval CE	Sel. CE	Eval CE	Δ eval CE	Sel. step	Q/K
VLM-MoE (decoder; continuation)	1.798	2.279	2.359	+0.561	34000	16
VLM-dense (decoder)	1.841	2.327	2.409	+0.567	16000	16
VLM-dense, output-KD only	1.841	2.327	2.408	+0.567	16000	16
VLM-dense, vision encoder only	1.841	1.855	1.853	+0.011	4000	24
VLM-dense, whole model	1.841	2.335	2.416	+0.575	16000	40
VLM-dense, dummy-image training (E1)	1.841	2.570	2.639	+0.798	15000	16
Plain LLM, official CE	1.378	2.447	2.513	+1.134	19000	16
Plain LLM, KD mixture	1.378	2.498	2.560	+1.182	17000	16

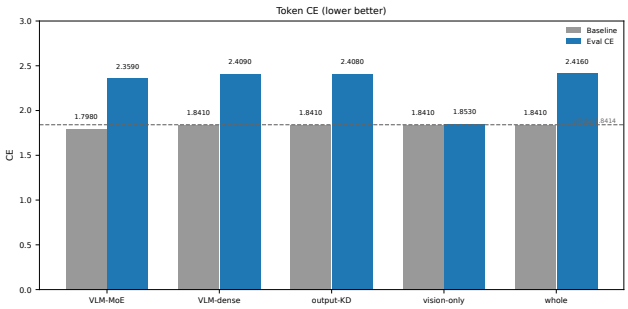


Figure M.1: Baseline (gray) versus post-DK eval (blue) token CE for the five principal VLM transfer arms (excluding the E1 dummy-image control; Table M.1); the dashed line is the official pretrained dense checkpoint (1.8414); CE is lower is better (↓).

Identity swap. Direct scores on unadapted dot weights degrade the language stack to CE 8.1 (PPL 3375) from the 1.841–1.798 baselines, so the swap alone is not usable; the Q/K transfer is required. The forward gate (Direct ≡ min-form softmax, max error 10^{-7}) was verified on every run.

Image-conditioned recovery (E1, E2, E3). In these single-seed, short-budget plain-LLM runs, the two supervision objectives give similar gaps (official CE +1.134, KD mixture +1.182), so we do not observe loss choice explaining the language-side gap. With identical loss and backbone the VLM gap is roughly half the LLM gap (+0.567 vs. +1.182). Three controls further describe this difference but do not isolate a mechanism. E1 (training-side): training with dummy image tokens early-stops with selection CE 2.570 and eval CE 2.639, versus 2.327 and 2.409 with image features. E2 (inference-side): withholding images from the image-trained model raises eval CE from 2.409 to 2.861, while feeding real images to the ablated model raises its CE from 2.639 to 2.718; the image-trained model depends on the image features at inference and the ablated model has no use for them. E3 (supervision channel): output-only KD (final-logit KD, no layer-wise attention matching) reaches essentially the same selection CE (2.327) and nearly identical eval CE (2.408) as the full KD mixture (2.409). In this run we observe no additional recovery from layer-wise attention matching, without excluding an effect at other budgets or seeds.

Whole-model replacement (E4). Replacing every dot attention in MiniMind-V—SigLIP2 encoder and decoder—composes from the arm results. The vision-encoder-only arm reaches selection CE 1.855; on the evaluation split, it reaches CE 1.853 versus baseline eval CE 1.841. Any configuration that includes the causal decoder self-attention (decoder-only 2.409, whole-model 2.416) keeps a gap of roughly +0.56 eval CE. In these single-seed, budgeted arms, the non-causal vision stack recovers better and the observed gap is concentrated in configurations containing causal decoder attention; the comparison does not attribute that gap to the mask or to any single mechanism.

Mechanism probes. Post-transfer probes (per layer, frozen checkpoints): centered-score Spearman 0.47–0.73, top-5 key overlap 0.47–0.77, attention KL 0.32–1.19, Direct entropy systematically higher by 0.5–1.5 nats. On the MoE model the frozen router redistributes: expert-routing agreement 0.66–0.77 (about 30% of tokens switch experts), and per-layer routing agreement tracks the attention KL (Pearson $\rho = 0.716$). Within a model, the per-layer correlation between dot attention entropy and post-transfer attention KL is weak (0.0–0.25, one negative value; Figure M.2); this weak association does not support a per-layer diffuseness gradient as the explanation for the modality difference, while input, architecture, and other factors remain unisolated. The ImageNet ViT-Small dot anchor measured on the 5,000-image geometry manifest has mean entropy 3.526 nats (range 2.821–4.071), a similar profile. In the short-budget text run, the KD mixture produces the closest attention geometry while the task gap is similar to that under the official loss; closer attention geometry is not accompanied by better logit recovery here, which does not establish a general mechanism.

From-scratch training. The same MiniMind-V stack (dense decoder + SigLIP2 encoder, both randomly initialized, every parameter trained) is trained from scratch on the SFT pool in four attention layouts: all-dot (baseline), Direct in the vision encoder only, Direct in the causal decoder only, and

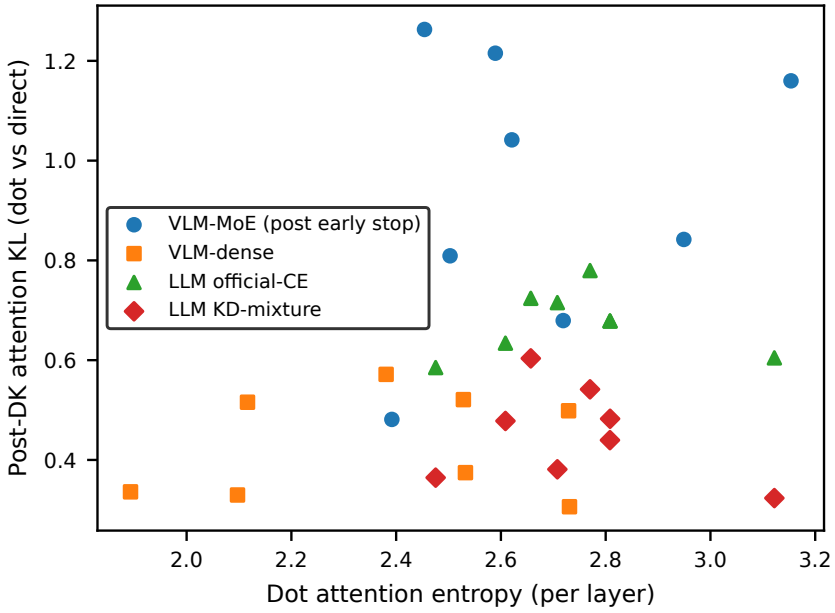


Figure M.2: Per-layer post-transfer attention KL (dot vs. direct) against the dot attention entropy, for the four transfer runs. The within-model correlation is weak (0.0–0.25, one negative); this single-seed probe does not support a per-layer diffuseness gradient as the explanation for the larger recovery in the image-conditioned runs, but it does not exclude other explanations.

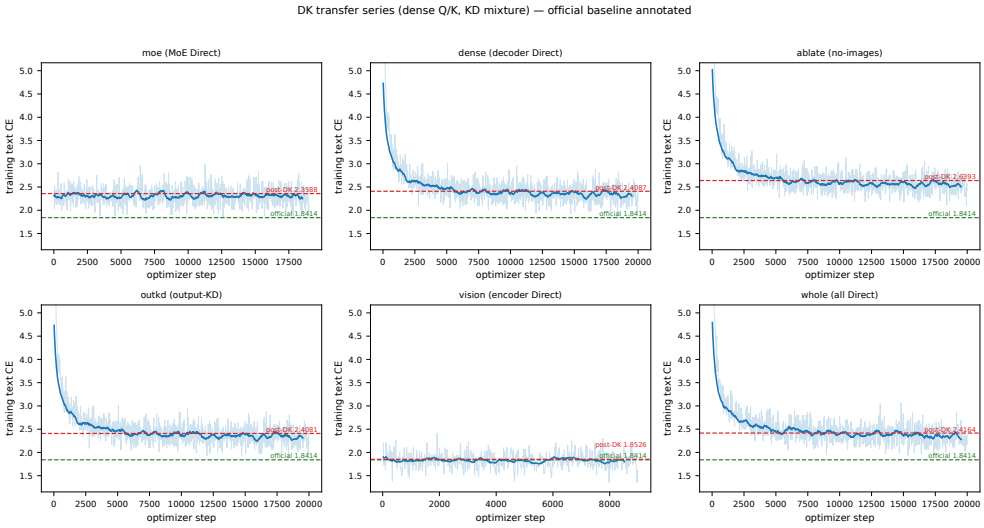


Figure M.3: DK transfer series: training-minibatch token CE against optimizer step for the six VLM arms (the official pretrained dense-checkpoint evaluation CE 1.8414 is the common green visual reference; the MoE arm's matching baseline is 1.798; each panel's red line is that arm's post-DK evaluation CE). The training-batch curves show relative optimization trajectories only. Same-split endpoint comparisons use the arm-specific baselines in Table M.1.

Direct everywhere (plain next-token CE, lr 5×10^{-4} cosine, warmup 500, batch 4×2, sequence 384, 10,000 steps, single seed). The attention-replacement percentage is measured on 200 eval rows both by layer count and by score-pair count ($B \cdot H \cdot T_q \cdot T_k$); because the 64 vision tokens are far shorter than the up-to-384-token text sequences, the vision encoder is 60% of the layers but only 5.9% of

the score computation, while the causal decoder is 40% of the layers but 94.1% of the score pairs (Table M.2).

Table M.2: From-scratch MiniMind-V (single seed, 10,000 steps, development evidence). "Layers" and "Score pairs" are the shares of attention computation performed by Direct (by layer count and by measured $B \cdot H \cdot T_q \cdot T_k$ pairs on the eval split). ROUGE-L is numerically similar across arms (0.287–0.296).

From-scratch arm	Layers	Score pairs	Sel. CE	Eval CE	PPL	BLEU-4
dot (baseline)	0%	0.0%	2.5146	2.5286	12.54	0.1740
vision Direct (full)	60.0%	5.9%	2.4394	2.4521	11.61	0.1745
decoder Direct (causal)	40.0%	94.1%	2.7024	2.7285	15.31	0.1601
all Direct	100.0%	100.0%	2.7122	2.7438	15.55	0.1616

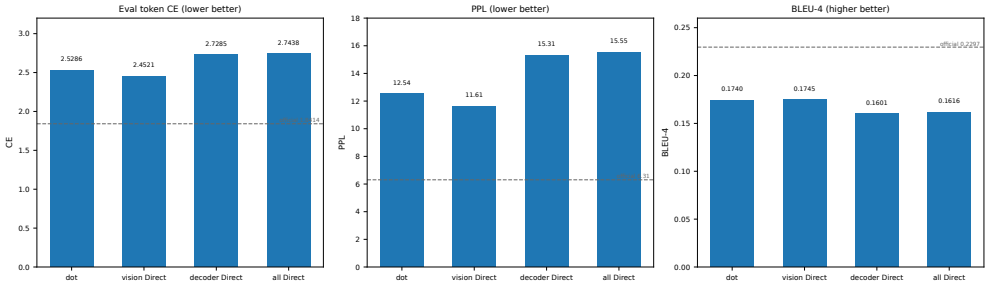


Figure M.4: Eval token CE, PPL, and BLEU-4 for the four from-scratch layouts (Table M.2); the dashed lines are the official pretrained checkpoint values (CE 1.8414, PPL 6.31, BLEU-4 0.2297); CE/PPL are lower is better, higher BLEU is better.

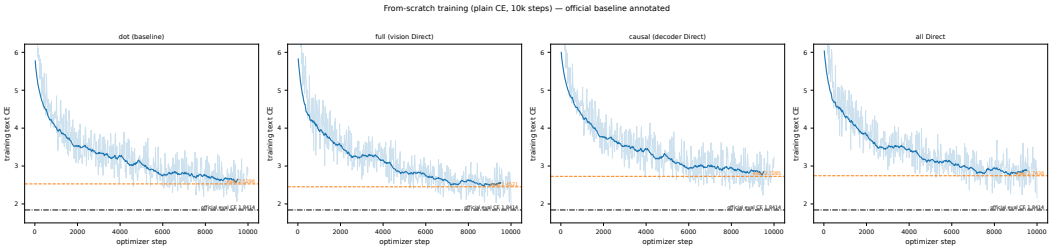


Figure M.5: From-scratch training: token CE against optimizer step for the four attention layouts (the official pretrained eval CE 1.8414 is the black dash-dot line; each panel's orange dashed line is that layout's post-training eval CE, Table M.2). All arms start at the same random-init CE; the vision-Direct arm tracks at or below the all-dot baseline, and the decoder-Direct arm tracks above it throughout. The all-Direct arm tracks above the dot baseline for most of training but its training CE falls below at the very end, while its held-out eval CE still finishes about $+0.2$ above dot's — a train-eval divergence (all-Direct overfits the training split harder).

From-scratch result. At equal single-seed budget, the vision-Direct arm matches or slightly exceeds the all-dot baseline on every eval metric (eval CE 2.4521 vs. 2.5286, PPL 11.61 vs. 12.54, BLEU-4 0.1745 vs. 0.1740), while both decoder-Direct arms finish about $+0.2$ CE behind (2.7285, 2.7438; Figure M.5). This is a development observation at one seed — no superiority claim. Its ordering (vision-Direct $\geq$ dot $>$ causal $\approx$ all) is consistent with the transfer runs and the translation arms of Section N; it suggests, but does not establish, that the tested full-attention vision stack is easier to adapt than the causal decoder under these budgets.

Distillation-free full-parameter finetuning (VLM). The same MiniMind-V stack is then finetuned from its DK/pretrained checkpoints with plain CE (no distillation), all parameters trainable, lr 5×10^{-5}

(warmup 500), batch 8, 16,000-step cap with selection every 1,000 steps. Two distillation-free controls start from the pretrained dot VLM (dense and MoE variants, 2.7978/2.9389); the six Direct arms start from their respective dense-transfer checkpoints (MoE, dense decoder, image-ablated, output-KD, vision-only, whole-model). Every arm ran the full 16,000 steps without early stopping.

Table M.3: Direct adaptation of MiniMind-V against the architecture-matched official pretrained checkpoints (2,000-row eval split; single seed, development evidence). "DK" is the dense-transfer result; "DK-FT" continues from the DK checkpoint with distillation-free full-parameter finetuning (16,000 steps, plain CE). The DK vision-only arm is numerically closest to the official dense checkpoint (1.8526 vs 1.8414); DK-FT improves the mid-pack arms by 0.3–0.6 CE, but no configuration bypasses its official checkpoint and FT slightly regresses the vision arm. For the MoE arm, the architecture-matched official baseline is 1.798 (Table M.1); DK/FT within-arm comparisons do not substitute the dense baseline for it.

MiniMind-V arm	DK eval CE ↓	DK-FT eval CE ↓	Δ (FT−DK)	Best
official pretrained dense dot (reference)	**1.8414**	—	—	—
official pretrained MoE dot (reference)	**1.798**	—	—	—
vision (encoder Direct)	**1.8526**	1.9050	+0.052	**DK**
moe (MoE Direct)	2.3588	2.0555	−0.303	FT
dense (decoder Direct)	2.4087	1.9959	−0.413	FT
outkd (output-KD)	2.4081	2.0264	−0.382	FT
whole (all Direct)	2.4164	2.0286	−0.388	FT
ablate (no-images)	2.6393	2.0292	−0.610	FT

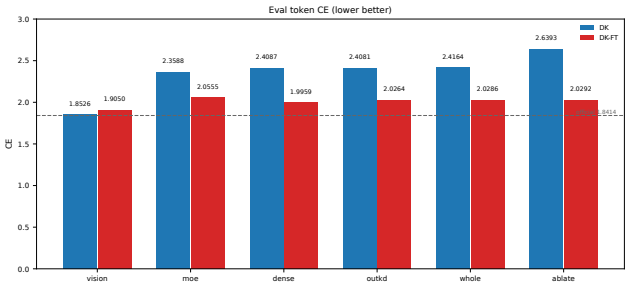


Figure M.6: DK (blue) versus distillation-free DK-FT (red) eval token CE per arm (Table M.3); the dashed line is the official pretrained dense-checkpoint reference (1.8414), while the MoE arm's architecture-matched official baseline is 1.798; lower CE is better (↓).

Result against the official checkpoint (VLM). Against the architecture-matched official pretrained checkpoints, dense transfer gives the best recovery among these runs (Figure M.3): the DK vision-only arm approaches the official dense-checkpoint value (eval CE 1.8526 vs 1.8414, PPL 6.38 vs 6.31, ROUGE-L 0.341 vs 0.343), and the DK decoder arms recover to 2.41–2.42. Distillation-free full-parameter finetuning from the DK checkpoints then improves the mid-pack arms by 0.3–0.6 CE (dense 2.4087→1.9959, whole 2.4164→2.0286, ablate 2.6393→2.0292; Table M.3) but *no configuration bypasses its architecture-matched official checkpoint*, and FT slightly regresses the best DK arm (vision 1.8526→1.9050; Figure M.7). The plain-CE controls trained from the official checkpoints degrade far below them (2.8150 and 2.9608), so they are not stronger references than the pretrained checkpoints. Within these single-seed runs, the tested FT continuation does not close the remaining gap; this is development evidence, not a general comparison of DK and FT.

Cross-series finding. Across the three training series—DK transfer, from-scratch, and distillation-free FT—the training curves show a repeated descriptive ordering: the vision-encoder Direct arm approaches the official dense-checkpoint value at DK time (Figure M.3), the causal-decoder arms plateau 0.5–0.8 CE above their corresponding baselines, and no FT configuration crosses its architecture-matched official reference (Figure M.7); the plain-CE controls converge highest of all.

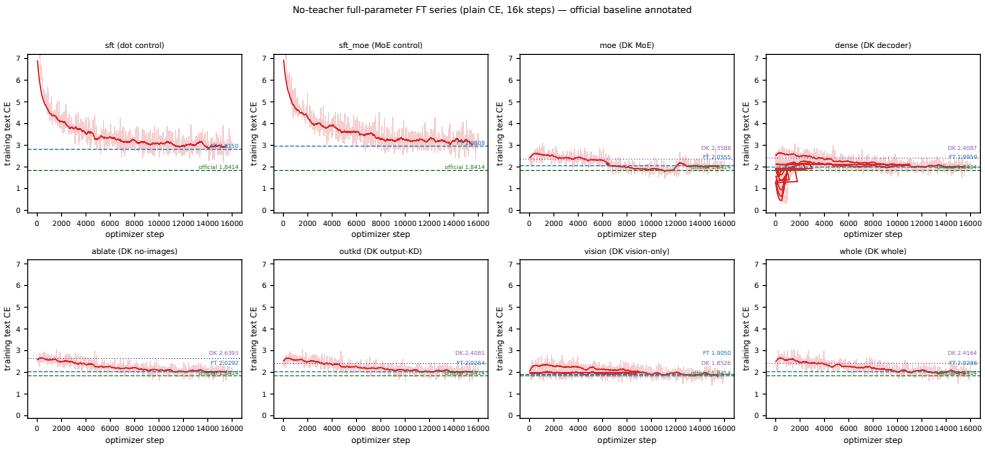


Figure M.7: Distillation-free full-parameter finetuning series: training token CE against optimizer step for all eight arms (green: official dense-checkpoint eval CE 1.8414; dotted: the arm's DK eval CE before FT; blue: the arm's post-FT eval CE). The architecture-matched official MoE baseline is 1.798. Every DK-initialized Direct arm trains down to 1.9–2.0 and its post-FT eval tracks the curve; the two controls converge far above (2.8–3.0), and none of the arms crosses its architecture-matched official reference.

N CAUSAL-MASK TEST ON A TRANSLATION MODEL

Model and data. The transfer study uses NLLB-200-distilled-600M (M2M100 architecture; 12 bidirectional encoder layers and 12 decoder layers containing causal self-attention and cross-attention; model width 1024, 16 heads, head width 64, sinusoidal positions, and additive masks). The Chinese–English corpus contains 24,752,392 sentence pairs; experiments use a 200,000-row training pool, a disjoint 2,000-row selection split, and 7,379 rows from WMT18/19/20 newstest for evaluation. All three attention types coexist in the same model, but their module roles and sequence statistics differ. The model and WMT Chinese–English corpus were obtained from ModelScope. Because WMT16 EN–DE was unavailable from the experiment host, the pre-execution-approved substitution used zh→en WMT18/19/20 newstest.

Protocol. Four arms replace encoder self-attention only (A1), cross-attention only (A2), causal decoder self-attention only (A3), or all three (A4). Each arm first records an identity swap, then optimizes only the Q/K projections with all other tensors frozen. The loss uses $\alpha = 0.9$, dense mix 0.5, output temperature $T = 3$ with T^2 scaling, learning rate 10^{-4} with cosine decay and 100 warm-up steps, gradient clipping at 1.0, bf16 autocast, selection every 1,000 steps, patience five with a minimum CE improvement of 5×10^{-3}, and a 3,000-step cap. A1 uses effective batch size 16, whereas A2–A4 use 12 because of memory limits. Optimizer-step budgets are therefore matched, but processed-example counts are not. Dynamic per-batch padding gives an average sequence length of about 53 under a maximum of 128. Trainable Q/K tensor counts are 24/24/24/72, and an integrity audit verifies that all other tensors remain identical. All cross-arm comparisons are treated as single-seed development evidence.

Metrics. Token CE (bf16, sum-reduction, padding masked to -100) and word-level BLEU-4, both sentence-averaged and corpus-aggregated (standard n-gram counts plus brevity penalty), on greedy decoding with the target language forced. The BLEU numbers are not comparable to published SACREBLEU scores (single reference, raw whitespace tokenization, greedy decoding); they are used for internal arm comparisons only. Qualitative spot checks produced fluent examples (for example "Tokyo is the largest city in the world."), which reduces but does not eliminate concern about a broken decoding pipeline; the absolute BLEU is used only for internal arm comparisons.

Results. Baseline (A0, dot, no training): token CE 1.6976, PPL 5.46, corpus BLEU-4 0.171 (1,000 generated sentences). Table N.1 reports every arm: identity-swap damage, selection CE, post-DK eval CE and BLEU, deltas, recovery percentages (100% = full return to the dot baseline), selected step, and integrity. None of the arms early-stopped within the 3,000-step cap.

Table N.1: Complete Stage-5 arm record (NLLB-200 zh→en; baseline CE 1.6976, corpus BLEU-4 0.171). Recovery of the identity damage: A1 81% (CE) and 75% (BLEU); A2 51% and 23%; A3 18% and -4%; A4 26% and 1%. "Ident" is the identity swap (Direct on unadapted dot weights); "Sel CE" the best selection-split token CE; "Post" the post-DK eval. Q/K is the number of changed Q/K tensors (integrity; 24/24/24/72). Selected steps: A1 2,000, A2 3,000, A3 2,000, A4 2,000. Arrows on the metric headers mark the direction in which the metric is better: ↓ lower is better (CE), ↑ higher is better (BLEU); read each arm against the A0 dot-baseline row.

Arm	Ident CE↓	Ident BLEU↑	Sel CE↓	Post CE↓	Post BLEU↑	ΔCE↓	ΔBLEU↑	Q/K
A0 dot baseline	—	—	2.5196	1.6976	0.171	—	—	—
A1 encoder	2.950	0.019	2.5958	1.9326	0.134	+0.235	−0.038	24
A2 cross	3.967	0.003	3.1994	2.8155	0.042	+1.118	−0.129	24
A3 decoder	5.404	0.059	4.5656	4.7352	0.054	+3.038	−0.117	24
A4 all	6.723	0.000	5.0891	5.3930	0.002	+3.695	−0.169	72

Decision. Under the prespecified descriptive rule, A1's recovery gap is at most half of A3's on both CE and BLEU (ratios 0.08 and 0.32). A2 meets that comparison on CE (0.37) but not BLEU (1.1). The observed ordering is therefore consistent with greater difficulty in the causal decoder: the encoder arm recovers 81% of the CE gap and 75% of the BLEU gap, cross-attention recovers 51% and 23%, and the decoder arm recovers 18% and none, respectively. Because module role, sequence

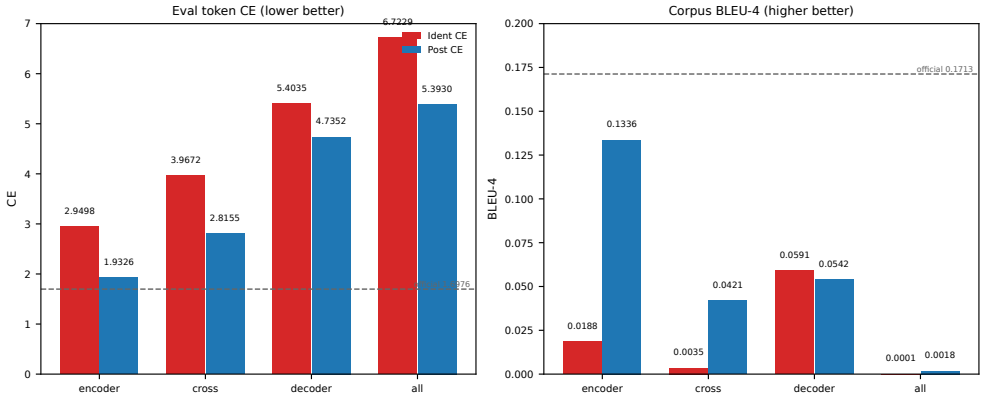


Figure N.1: Grouped bars for the Stage-5 arm record (Table N.1): identity-swap damage (red) versus post-DK eval (blue) token CE (left) and corpus BLEU-4 (right) per arm; the dashed line is the dot baseline (CE 1.6976, BLEU 0.171). lower CE is better (↓), higher BLEU is better (↑), read against the dashed baseline.

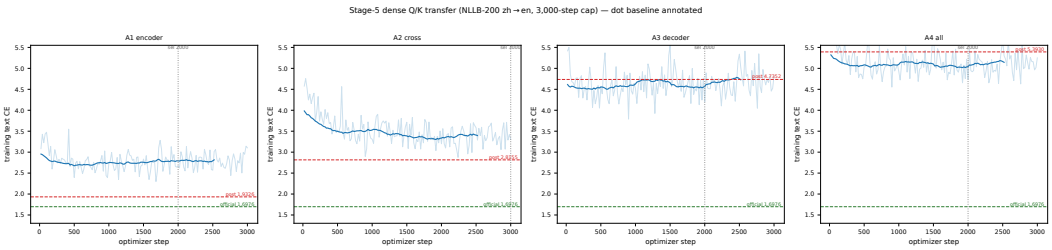


Figure N.2: Training series for the four Stage-5 dense Q/K transfer arms (training-batch token CE against optimizer step, 3,000-step cap). The green line is the dot baseline eval CE 1.6976; each panel's red line is that arm's post-DK newstest eval CE; the gray dotted line marks the selected checkpoint step (A1 2,000, A2 3,000, A3 2,000, A4 2,000). Every arm stays above the baseline; recovery of the identity-swap damage is reported in Table N.1.

distribution, effective sample count, and optimization difficulty are not controlled, this rule does not isolate the causal mask as the cause.

Interpretation. The observed ordering is consistent with greater difficulty in arms containing causal decoder attention. However, language pair, module role, sequence statistics, batch size, processed-example count, and optimization difficulty are uncontrolled, so masking is not identified as the cause. MiniMind-V shows a similar descriptive association (Appendix M): its non-causal SigLIP-encoder arm approaches the dot baseline, whereas configurations containing causal decoder attention retain a roughly $+0.56$ CE gap. This cross-model resemblance is not a mechanism claim.

Follow-up result. Relaxing the KD weight toward the label CE moves further on the in-domain selection split—selection CE 2.2256 and 2.0081 versus the dot baseline's 2.5196—so the Direct encoder with a CE-dominated objective *surpasses* dot on the selection split (single-seed observation)—but the held-out newstest gap bottoms out at $+0.135$ CE (1.8331, $\alpha = 0.5$); near-pure CE ($\alpha = 0.1$) overfits the training domain (selection 2.0081 but test 1.8859). Combining both non-causal stacks (encoder + cross, 48 tensors) does not help: the cross-attention component recovers poorly (test CE 2.8007, BLEU 0.049), so the encoder-only arm remains the best non-causal configuration. Within the tested single-seed Q/K-only budget, the dot baseline is approached but not reached on the held-out test set; whether longer Q/K-only training or adaptation beyond the Q/K projections can reach or exceed it remains untested.

Table N.2: Follow-up on the bidirectional (encoder) arm. All Direct rows are single-seed development evidence with Q/K-only integrity (24 or 48 tensors changed, final equals selected). Selection CE is on the 2,000-row in-domain split; test CE and BLEU on the zh2en newstest sets. Arrows on the metric headers mark the direction in which the metric is better: ↓ lower is better (CE), ↑ higher is better (BLEU); read each row against the dot-baseline row.

Encoder-arm follow-up	Selection CE↓	Test CE↓	Δ CE↓	Corpus BLEU↑
dot baseline	2.5196	1.6976	—	0.171
A1 α=0.9, initial 3k	2.5958	1.9326	+0.235	0.134
A1 α=0.9 continuation (retained 8k; ended 17k)	2.5352	1.8523	+0.155	0.143
A1 α=0.5, 12k steps	2.2256	1.8331	+0.135	0.146
A1 α=0.1, 12k steps	2.0081	1.8859	+0.188	0.138
encoder + cross, 6k steps	3.0168	2.8007	+1.103	0.049

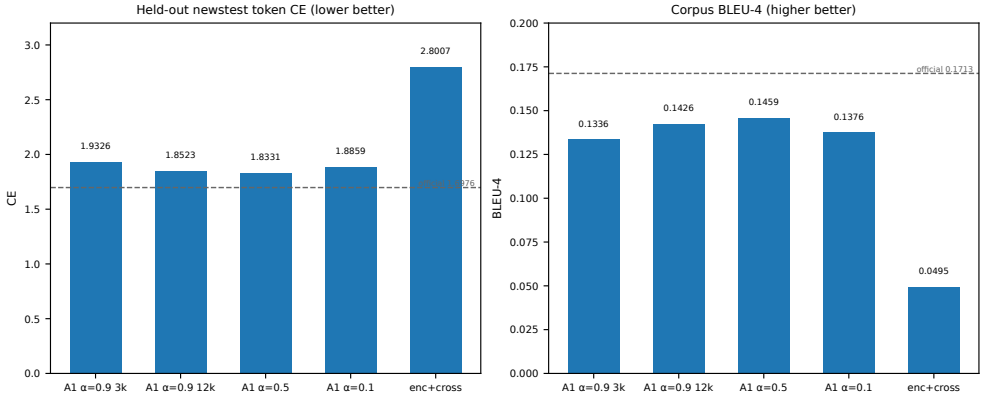


Figure N.3: Held-out newstest token CE (left) and corpus BLEU-4 (right) for the five encoder-arm follow-ups (Table N.2); the dashed line is the dot baseline (CE 1.6976, BLEU 0.171). lower CE is better (↓), BLEU higher is better (↑).

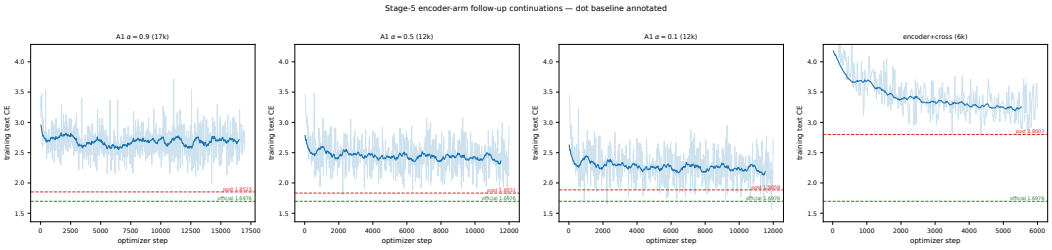


Figure N.4: Follow-up training series on the bidirectional encoder arm (training-batch token CE against optimizer step): A1 continued to 17,000 steps at $\alpha = 0.9$, relaxed-KD runs at $\alpha = 0.5$ and $\alpha = 0.1$ ran 12,000 steps each, and the encoder+cross combination ran 6,000 steps. The green line is the dot baseline eval CE 1.6976; each panel's red line is that run's held-out newstest eval CE (Table N.2). Only the $\alpha = 0.5$ run approaches the baseline (1.8331) without reaching it; $\alpha = 0.1$ overfits the training domain.

Distillation-free full-parameter finetuning (NLLB). The Stage-5 protocol is repeated without distillation: the four stage-5 arms plus the non-causal encoder+cross combination are finetuned with plain CE, all parameters trainable, lr 5×10^{-5} (warmup 500), batch 12, 16,000-step cap, selection every 1,000 steps, starting from their dense-transfer (DK) checkpoints; the dot control starts from the pretrained NLLB checkpoint.

Table N.3: Direct adaptation of NLLB-200 (zh→en, newstest held-out) against the single reference, the official pretrained checkpoint. "DK" is the dense-transfer initializer (3,000-step cap for encoder and cross; 6,000 steps for encoder+cross); "FT" continues each listed Direct row from its corresponding DK checkpoint with distillation-free full-parameter finetuning (16,000 steps, plain CE). The DK encoder arm is the strongest Direct DK configuration among the rows shown here (1.9326 vs 1.6976); the separate $\alpha = 0.5$ Q/K follow-up reaches 1.8331 and appears in Table N.2. FT does not improve the encoder, and the causal-decoder arms collapse under the push-limit protocol (Table N.4). Single seed, development evidence.

NLLB-200 arm	DK CE ↓	DK BLEU ↑	FT CE ↓	FT BLEU ↑
official pretrained dot (baseline)	**1.6976**	**0.1713**	—	—
encoder	**1.9326**	**0.1336**	2.3826	0.1224
cross	2.8155	0.0421	**2.6173**	**0.1200**
noncausal (encoder+cross)	2.8007	0.0495	**2.6665**	**0.1045**
dot (FT control)	—	—	2.4066	0.1278

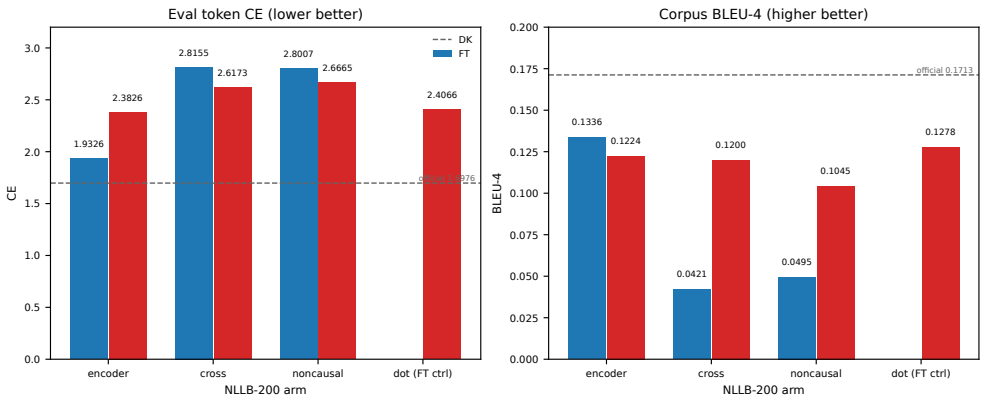


Figure N.5: Grouped bars for the NLLB-200 DK/FT result table (Table N.3): DK (blue; 3,000 steps for encoder and cross, 6,000 for encoder+cross) and FT (red, 16,000-step distillation-free finetuning) held-out eval CE (left panel) and corpus BLEU-4 (right panel) per arm; the dot control has no DK bar, and the official pretrained checkpoint is the dashed line (CE 1.6976, BLEU 0.171). lower CE is better (↓) and higher BLEU is better (↑), read against the dashed baseline.

Table N.4: Push-limit protocol outcome (full 16,000-step cap, no early stopping): the two arms containing Direct in the causal decoder. The selection loss is driven to near zero and greedy decoding collapses (corpus BLEU 0.000), so the held-out numbers are reported separately rather than as comparison rows. Single seed, development evidence.

Push-limit arm	Sel. CE ↓	Eval CE ↓	Corpus BLEU-4 ↑
decoder (causal)	0.0772	0.8527	0.0000
all	0.1209	1.1823	0.0000

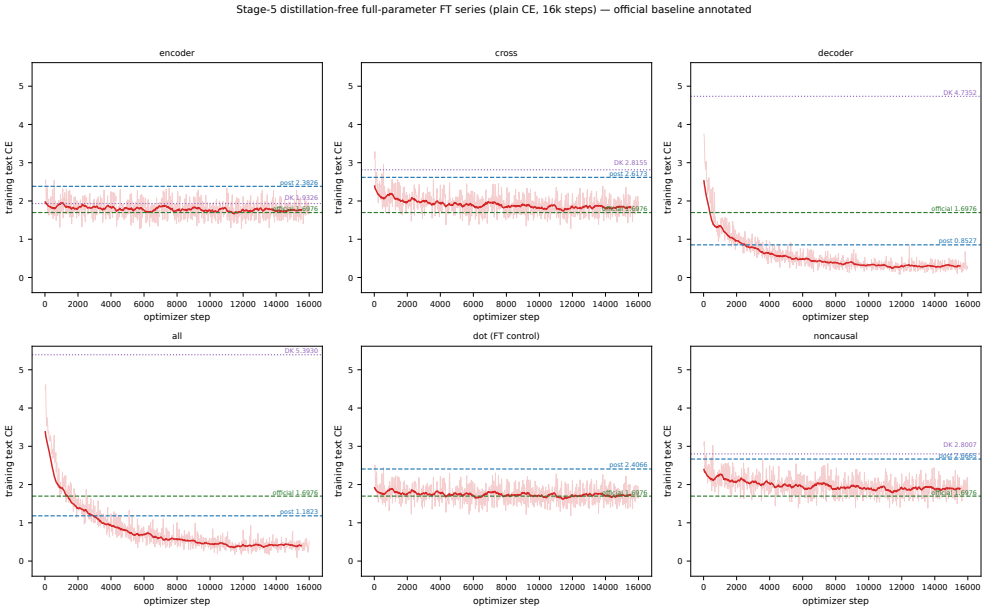


Figure N.6: Distillation-free full-parameter finetuning series (training-batch token CE against optimizer step, plain CE, 16,000 steps) starting from each arm's DK checkpoint; the dot control starts from the pretrained checkpoint. The green line is the official baseline CE 1.6976; dotted purple lines are the arms' DK eval CEs before FT; blue lines are the post-FT eval CEs (Table N.3). The causal-decoder arms (decoder, all) collapse toward zero CE while their corpus BLEU drops to 0.000 (Table N.4).

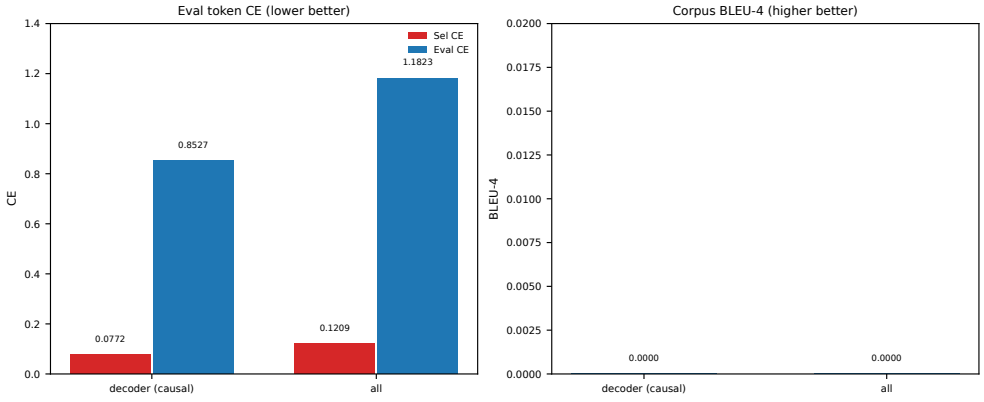


Figure N.7: Push-limit bars (Table N.4): selection CE (red) versus held-out eval CE (blue) per causal-decoder arm, and the collapsed corpus BLEU-4 (both 0.000). lower CE is better (↓), higher BLEU is better (↑).

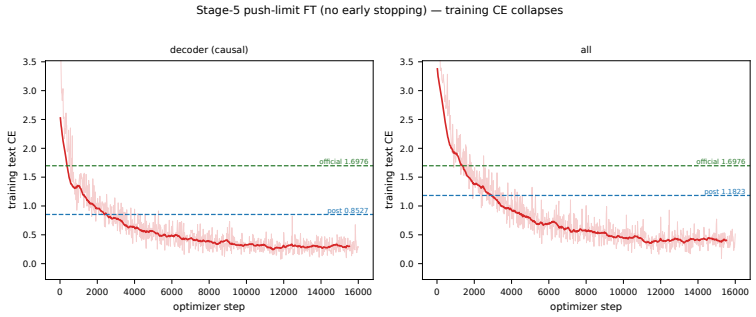


Figure N.8: Push-limit finetuning series for the two causal-decoder arms (full 16,000-step cap, no early stopping). Training CE collapses toward zero (selection CE 0.0772 and 0.1209), and greedy decoding fails (corpus BLEU 0.000), so the held-out eval CEs (0.8527 and 1.1823) are reported separately in Table N.4 rather than as comparison rows.

Result against the official checkpoint (NLLB). With the official pretrained model as the single reference, the 3,000-step DK encoder is the strongest Direct row among those shown in Table N.3: it reaches 1.9326 CE / 0.1336 BLEU versus the official 1.6976 / 0.1713. The separate $\alpha = 0.5$ Q/K follow-up is the best held-out Direct NLLB result overall (1.8331 CE; Table N.2), but it is not one of the rows in the DK/FT table and remains above the baseline. Distillation-free FT does not improve on the encoder row: the encoder regresses under FT (2.3826 / 0.1224); cross and noncausal improve over their own DK results but stay far behind the official; the dot FT control (2.4066 / 0.1278) is itself worse than the untrained official. The causal-decoder arms drive the selection loss to near zero under the push-limit protocol and collapse under greedy decoding (BLEU 0.000; Table N.4). Under their respective distillation-free full-parameter protocols, the causal-decoder NLLB arms collapse, whereas several VLM DK-FT arms improve over their own DK checkpoints; these cross-task outcomes are not directly comparable. As with all distillation-free results here, these are single-seed development observations.

Caveats. Each arm uses one seed. The 3,000-step cap may under-train the harder A3/A4 arms, and cross-attention's BLEU recovery lags its CE recovery. The evidence supports better recovery for the tested encoder arm than for the tested causal decoder arm; it does not isolate causal masking from module role, data distribution, effective sample count, or optimization difficulty.

Longer encoder-arm follow-up. Continuing A1 to 12,000 steps with the same objective reduces its CE gap from $+0.235$ to $+0.155$ and its BLEU gap from -0.038 to -0.029. A subsequent learning-rate-restart segment ended by early stopping at global step 17,000 and retained the earlier global step 8,000, without further held-out improvement (Table N.2).

O HARDWARE ESTIMATION AND SYNTHESIS STUDY

Illustrative cost model. The hardware estimates use an illustrative fixed-point static-CMOS area model. A two-input NAND is counted as four transistors and a one-bit full adder as 28. Direct uses one two's-complement subtraction and a sign threshold per coordinate, followed by one score-level negation after accumulation.

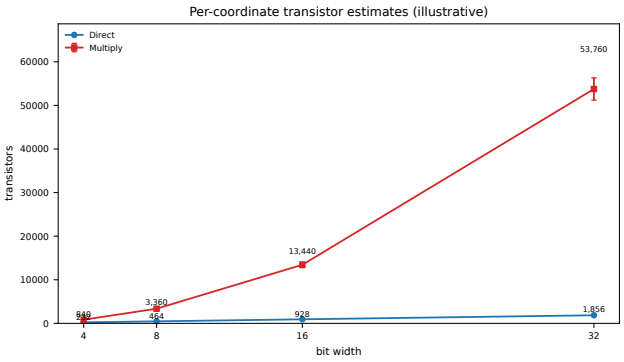


Figure O.1: Illustrative per-coordinate transistor estimates at 4/8/16/32 bits: the Direct subtraction/threshold model grows linearly (232/464/928/1856) while the multiplier model grows quadratically (range bars at 800–880 / 3200–3520 / 12800–14080 / 51200–56320). A two-input NAND counts as 4 transistors and a one-bit full adder as 28; the common accumulation tree, softmax, AV, projections, and MLPs are excluded; lower area is better (↓).

Table O.1: Illustrative fixed-point score-arithmetic estimates at 8, 16, and 32 bits. Per-coordinate transistor counts compare the linear-width Direct subtraction/threshold model with a quadratic-width multiplier model under stated NAND and full-adder assumptions. The estimates exclude memory, routing, accumulation, softmax, value aggregation, projections, and MLPs and are not synthesis results.

Item	Direct estimate	Multiply estimate	Ratio / treatment	Scope / limitation
8-bit per-coordinate estimate	464	3200–3520	6.9–7.6x	illustrative fixed-point area model
16-bit per-coordinate estimate	928	12800–14080	13.8–15.2x	illustrative fixed-point area model
32-bit per-coordinate estimate	1856	51200–56320	27.6–30.3x	illustrative fixed-point area model
NAND2 / full-adder assumptions	4 / 28 transistors	–	model inputs	library-dependent
score-level negation	one bit-width-linear operation	–	counted separately	not per coordinate
common accumulation tree and remaining model	–	–	excluded/common	AV, projections, MLP still multiply
relationship to implementation study	–	–	separate evidence	synthesized score units reported below

The analytic model motivates, but does not validate, efficiency. Across this illustrative fixed-point static-CMOS model, the per-coordinate Direct estimate grows linearly with bit width, whereas the multiplier estimate grows quadratically, yielding larger modeled ratios at 4, 8, 16, and 32 bits. These counts omit wires, SRAM, clocking, control, pipelining, softmax, AV, projections, and MLP layers, and the common accumulation tree is not a Direct-specific saving. We therefore use this model only as motivation. The following standard-cell study separately reports mapped area and functional cycle counts, while distinguishing historical timing and activity records from admitted results. Neither analysis establishes measured silicon or end-to-end accelerator efficiency. Linearity here is in operand bit width, not sequence length; dense scoring retains $\Theta(n^2 d_k)$ coordinate operations.

Target and modeled designs. Two static (fixed-schedule, non-programmable) ViT-Small/16 inference-engine designs are compared: a *standard design* whose every multiply is represented by a signed MAC cell (DOT_PE), including the score stage, and a *Direct design* that is identical except the score stage uses the multiplication-free Direct cell (DIRECT_PE: subtract $\rightarrow$ clamp-min-0 $\rightarrow$ negate $\rightarrow$ accumulate). Both modeled designs assume all 22M weights on-chip (weight SRAM and activation SRAM are identical by construction). The modeled datapath has 128 GEMM MAC lanes, six score units (16 lanes each, 64 coordinates, four cycles per score), a shared softmax unit, and single-port SRAM at 256 bits/cycle. Trace generation uses the frozen Dot reference under the generic preprocessing and a deterministic 100-image manifest; no Direct-checkpoint or across-seed hardware variation is claimed.

Matched integer contract. For width W, both synthesized score units accept signed two's-complement W-bit Q/K inputs and expose the same signed $2W$-bit accumulator and score output (32 bits at the primary $W = 16$ point). Accumulation wraps modulo 2^{2W} after every coordinate; neither unit saturates. Dot forms an exact $2W$-bit product before the wrapped addition. Direct forms the subtraction in $W + 2$ bits, clamps it at zero, and performs the same wrapped accumulation. A post-scale is outside these synthesized score cores. The model-level quantization proxy retained during development instead used per-forward-batch max-absolute scaling and wide integer accumulation. Because it does not instantiate this RTL contract, its accuracy values are excluded from the paper's hardware evidence. The bit-exact gate below validates the stated RTL arithmetic.

Functional simulation (bit-exact gate). The cells are bit-exact against the Python fixed-point reference (FIXEDPOINT.PY, the specification): for every (operator, width) $\in$ {dot, direct} $\times$ {4, 8, 16, 32}, 2,007 edge/random vectors and 57,600 real ImageNet Q/K trace vectors (extracted from the frozen Dot reference under the generic preprocessing on 100 images $\times$ 12 layers) match the reference exactly; the shared softmax unit matches its LUT reference; and the six-head score tile for one encoder layer completes $197 \times 197 \times 4 = 155236$ cycles exactly in RTL simulation. This cycle-exact result applies only to the score tile. The other stage counts, the per-layer subtotal, and the whole-model total below are entries in an analytical static schedule, not cycle-exact whole-model RTL measurements. The gate passes before any synthesis.

Mapped area and source provenance. The retained synthesis statistics give 180135.3 versus 20784.9 μm^2 for the 16-bit Dot and Direct score units, respectively (Dot/Direct $= 8.67\times$). Their cell counts agree with the corresponding placed designs. The recorded width sweep gives Dot/Direct ratios of 2.58, 4.75, and 16.2 at 4, 8, and 32 bits. This trend is consistent with the illustrative quadratic partial-product and linear-width subtract/rectify models; mapped area also includes accumulation and control. The 16-bit single-cell areas are 11140.7 and 1352.5 μm^2. The scripts select Sky130 HD TT at 25 °C and 1.8 V; the exact historical container, library bytes, and complete producing-run versions are not fully authenticated. Tool and library provenance is Yosys, OpenROAD, and the SkyWater SKY130 PDK (Wolf & Glaser, 2013; Ajayi et al., 2019; SkyWater PDK Authors, 2020).

An independent static check of the hash-authenticated, currently retained 16-bit netlists reproduces the area ratio without rerunning synthesis. It counts all 23,942 Dot and 2,756 Direct instances under the same authenticated TT library. Every referenced cell's Liberty area equals its LEF width–height product, giving total mapped-cell footprints of 180,135.2640 and 20,784.9344 μm^2, respectively ($8.6666\times$). The area-unit interpretation is cross-checked against the matching OpenROAD flow, not an explicit Liberty area-unit declaration. This check substantiates the current-netlist mapped-area comparison; it does not establish that these bytes produced the historical reports, nor validate post-clock-tree or routed area, timing, power, or whole-chip savings.

Historical timing records, excluded from performance claims. The retained placement-parasitic reports contain worst setup slacks of -1.32 and +1.26 ns under a 100 MHz constraint. The historical frequency estimates were calculated as reciprocals of the implied path delays, rather than verified by closure runs at those frequencies. The recovered flow creates the clock after its clock-tree-synthesis command; the retained paths show zero clock-network delay and the power reports show zero Clock-group power. These artifacts therefore do not demonstrate a clock-tree implementation. Moreover, the surviving later physical-flow logs are failed reruns, not the producing transcript for the retained placement and timing reports. We preserve these values in Table O.2 as historical diagnostics, not demonstrated $F_{\max}$, timing closure, or a common feasible operating clock.

Historical activity and power, excluded from energy claims. The activity test used 2,000 vectors per unit: 500 edge/random vectors and 1,500 shared frozen-Dot Q/K trace vectors, with a recorded duration of 80,030 ns. It was not a population of 2,000 model-trace vectors. The retained converter assigns a bus-wide transition count to every bit, reverses descending-bus dwell assignments, omits final dwell and constant ports, and assumes a timestamp unit rather than reading the VCD time scale. The two SAIF files agree on all 514 non-clock input records, but equality of incorrect records does not validate their activity. They contain only top-level port activity; internal activity is propagated rather than fully annotated from gate-net traces. In addition, the reported trace-power values are carried forward by the summary aggregator without an authenticated producing power report. We consequently exclude these power values, their ratios, and derived energy, capacitance, hotspot, and thermal quantities from the quantitative evidence. The numerical records are retained only to make the exclusion auditable; they do not predict a corrected power ratio.

Table O.2: Historical records excluded from hardware performance claims. The slack and reciprocal-delay calculations describe an ideal-clock placement analysis with incomplete producing-run provenance. The power row is an unvalidated legacy summary: its producing report is missing and the available activity conversion is defective. It is not a corrected measurement or estimate.

Historical record	Dot	Direct
Recorded slack at 100 MHz (ns)	-1.32	+1.26
Reciprocal-delay calculation (MHz; not verified $F_{\max}$)	88.3	114.4
Legacy power summary (W; excluded)	0.894	0.079

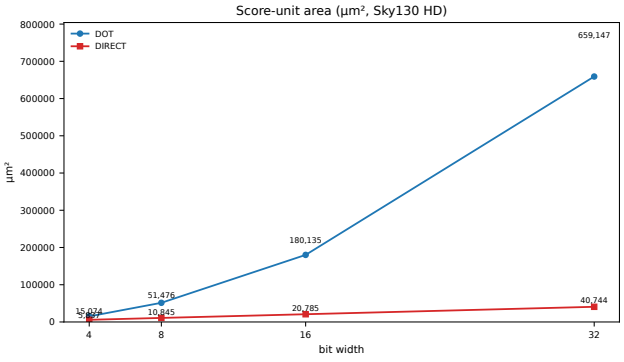


Figure O.2: Synthesized score-stage unit areas (μm^2, Sky130 HD) for the multiplier-based (blue) and Direct (red) score units at 4/8/16/32 bits; the recorded areas follow the quadratic-width multiplier and linear-width subtract/rectify model trends, with Direct smaller at every tested width (8.67$\times$ at 16 bits; 4.75$\times$ at 8 bits, 16.2$\times$ at 32 bits, 2.58$\times$ at 4 bits). The 16-bit MAC cells are 11140.7 versus 1352.5 μm^2; lower area is better (↓).

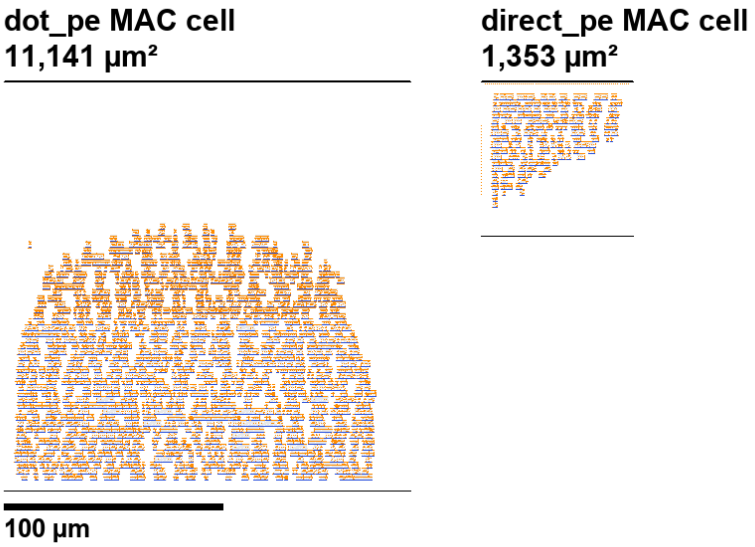


Figure O.3: Placed-only Sky130 HD layouts of the two 16-bit MAC cells at a common scale (2.2 px/µm, 100 µm scale bar): the multiplier-based dot_pe cell (11,141 μm^2) is 8.2× larger than the multiplication-free direct_pe cell (1,353 μm^2), consistent with the quadratic-width partial-product and linear-width subtract/rectify models. The score units replicate these MAC lanes, so the cell pair conveys the area advantage without repeating the replicated tile; routing is omitted for readability. Area is lower is better.



Figure O.4: Placed-only Sky130 HD layouts of the two full 16-bit score units at a common scale (2.2 px/µm, 100 µm scale bar): the multiplier-based DOT score unit (180,135 μm^2) is 8.67× larger than the multiplication-free DIRECT score unit (20,785 μm^2). The score units integrate the MAC datapaths plus accumulation and control logic; routing is omitted for readability. Area is lower is better (↓).

Full-system area ledger (16-bit, assuming all weights on-chip). Logic: 128 × DOT_PE GEMM arrays (1.43 mm^2), the score stage (6 × score unit: 1.08 mm^2 for the Dot design vs. 0.13 mm^2 for the Direct design), shared softmax (1.85 mm^2, modeled; its full synthesis stalls the tool) and LN/GELU (0.24 mm^2, modeled). SRAM: weight 22M × 16 bit (222.6 mm^2) and activation 2 × 197×384×16 bit (1.5 mm^2), 6T + 15% periphery at the library-documented 0.55 μm^2/bit. Modeled totals are 228.772 mm^2 (2.450 G transistors) for the Dot design and 227.816 mm^2 (2.449 G) for the Direct design; the identical weight SRAM makes the modeled SRAM-inclusive total-area ratio 1.0042, and the score stage is 0.47% (Dot) / 0.05% (Direct) of the total area. This SRAM-inclusive ledger is not a full-chip synthesis, place-and-route result, or silicon measurement.

Table O.3: Analytical static full-forward schedule for both designs. Only the 155236-cycle score-tile row is a cycle-exact RTL result; the other stage counts and every subtotal or extrapolation are analytical ledger entries. The designs have the same modeled count because the score stage changes only the cell type (same lanes/protocol). No operating frequency or end-to-end latency is inferred from these cycle counts.

Static-schedule term	Cycles	Basis
QKV GEMM (197×384×1152), per layer	680832	analytical
Six-head score tile, per layer	155236	cycle-exact RTL
Shared softmax (6 × 197 rows), per layer	1182	analytical
AV GEMM (6 × 197×197×64), per layer	117018	analytical
Output projection, per layer	226944	analytical
MLP (384→1536→384), per layer	1815552	analytical
LN/GELU, per layer	7880	analytical
Per-layer subtotal	3004644	analytical sum
Encoder stack (12 layers)	36055728	12× subtotal
Patch embedding (once per image)	451584	analytical
Pipeline fill	1392	analytical
Total per image	36508704	analytical ledger

Logic-only comparison (SRAM excluded). Within the analytical ledger and excluding the identical SRAM, the Dot design's modeled logic is 1.2623× the Direct design's (4.60 versus 3.64 mm^2, 4.90M versus 3.88M transistors): the entire difference (0.96 mm^2) is the score stage (6 units, 8.67× per unit), which is 23.5% of the Dot design's modeled logic versus 3.4% of the Direct design's. All other modeled logic (GEMM arrays, shared softmax, LN/GELU) is identical in the two designs.

Schedule and timing scope. The analytical ledger sums to 3004644 cycles per encoder layer, hence $12 \times 3004644 = 36055728$ cycles for the encoder stack. Adding 451584 patch-embedding cycles and 1392 pipeline-fill cycles gives 36508704 modeled cycles per image for both designs. Only the 155236-cycle score-tile entry was observed cycle-exactly in RTL simulation; the rest is analytical. These cycle counts do not establish a duration: neither a common feasible score-unit clock nor a whole-chip operating clock has been verified. Equal modeled schedules are not evidence of an end-to-end speedup.

Claim scope. The bit-exact functional checks pass at 4, 8, 16, and 32 bits, and the mapped 16-bit score-unit area favors Direct by 8.67×, with the same direction in the recorded width sweep. The analytical schedules are identical in cycles, and the SRAM-inclusive area ledger changes by only 0.42%. The historical timing and power records are not used to establish frequency, latency, power, or energy advantages. The supported implementation result is a flow-specific reduction in mapped score-unit area, separate from the floating-point checkpoint accuracy study.

P Adapting Causal Attention without a Mask: the Mirror-Direct Method and Its Ablations

Method: the mirror-Direct future-branch attention. Direct attention computes the score $-\sum_r \max(0, q_r - k_r)$ per query–key pair; the operator is asymmetric in its two arguments, and flipping the subtraction direction yields the mirror-Direct score $-\sum_r \max(0, k_r - q_r)$. Both share the same subtract–rectify–accumulate datapath; only the operand order changes. The proposed method uses this asymmetry to replace the causal mask: instead of applying $-\infty$ to the upper (future) triangle of the attention matrix, the future positions are scored with the mirror-Direct operator, while the causal positions keep the standard Direct scores. No $-\infty$ is ever applied, so the future participates in every forward pass — but through a direction-flipped geometry.

Protocol. The dense-transfer protocol is unchanged from Section N and Appendix M: only the 16 decoder Q/K projections are trained (KD mixture, $\alpha = 0.9$, dense-mix 0.5, output temperature $T = 3$), and the attention-matching KL anchors only the causal part ($s \leq t$) of the adapted attention to the dot reference ("causal_only"), so the future branch receives gradients only through the output-KD and CE terms. All arms run 6,000 steps with selection every 1,000 under the crash-resilient save-eval-continue protocol.

Identity swap and metrics. On the unadapted dot checkpoint the mirror-Direct future branch performs poorly (identity swap: eval CE 10.52, PPL 36,919, AR BLEU 0.001), consistent with a mismatch between causal-mask-pretrained weights and unadapted future scores. The single-seed, 6,000-step DK then recovers 97.4% of the measured CE damage: eval CE 2.0637, AR BLEU 0.1732 (Table P.1). Throughout, we report two complementary numbers: the full-context eval CE (future present) measures modeling quality, and the AR BLEU-4 (greedy left-to-right decoding, future absent) measures the generation the deployment cares about; their difference is reported as the pairing cost of future participation.

Ablation step 1: removing the mask alone (dot). Before evaluating the method, we isolate what removing the mask alone does, keeping the future triangle at the operator's plain scores. For the *dot* operator this arm (unmasked dot) reaches eval CE 1.8168 and AR BLEU 0.2346— both numerically better than the official masked reference (1.8414 / 0.2297). Two observations are consistent with this result: the causal_only anchor constrains past attention toward the dot reference, and the future branch receives no attention-level supervision. The comparison is leakage affected because full-context CE exposes future tokens and is not directly comparable with masked causal CE. A self-distillation control isolates this interpretation within this run: applying the identical DK protocol to the dot operator with the causal mask intact (masked dot self-distillation, Table P.1) recovers essentially nothing—eval CE 1.8412 against the official 1.8414, with the selection trajectory flat at $\approx$1.839 from step 3,000 onward. This makes the unmasked dot's CE gain consistent with the mask change rather than an extra-distillation gain under this single-seed, 6,000-step protocol. The control also recalibrates the AR side: the masked self-distillation shows the same +0.003 AR BLEU-4 over the official as the unmasked dot, so that apparent generation gain is DK-adaptation noise, not a mask effect. The qualifier "in this protocol" matters: unanchored longer training could increase future exploitation and degrade AR generation; the NLLB distillation-free FT collapse motivates this risk but is not a controlled prediction for this model.

Ablation step 2: removing the mask alone (Direct). For the *Direct* operator the same ablation inverts the picture: the full-context eval CE 2.3790 is slightly *below* the masked Direct reference (2.4087) because the unmasked row includes future information (Table P.1); deployable autoregressive BLEU falls to 0.0687 from the masked value 0.1897. Direct's subtract–rectify scores are flatter and higher-entropy than dot scores in these checkpoints, and the model appears to rely on future content that is unavailable at left-to-right decoding. This is an interpretation of one development run, not an isolated causal mechanism.

Ablation step 3: the mirror direction (the proposed method). Adding the mirror-Direct direction to the future triangle changes both metrics relative to unmasked Direct: eval CE drops from 2.3790 to 2.0637 and AR BLEU rises from 0.0687 to 0.1732. One interpretation is that the flipped subtraction makes future content less directly exploitable by the Q/K learned for past positions. In this single-seed,

6,000-step development run, the resulting endpoint combines lower full-context CE than the masked decoder or unmasked Direct with substantially higher AR BLEU than unmasked Direct. Thus, among the modifications tested under this budget, the mirror direction is the only one that improves both reported metrics; the experiment does not isolate the proposed mechanism beyond this setup.

The pairing cost. The mirror arm is not free: it still sits $+0.23$ CE and -0.057 AR BLEU above the official masked reference. We call this residual gap the pairing cost of future participation: full-context evaluation permits future use, whereas AR decoding removes it. This single-seed, 6,000-step comparison does not isolate how much of the gap is caused by reliance on future content. Reporting both metrics prevents the full-context CE alone from overstating the method's quality.

Table P.1: Unmasked-attention ablation with the mask-pretrained Q/K init (single seed, 6,000 steps, development evidence). Full-context CE is the reference-forced CE: masked rows are future-absent by construction (the causal mask), unmasked rows are future-present, so unmasked rows carry a future-leakage advantage in this column; AR BLEU-4 is the BLEU-4 of greedy autoregressive (left-to-right) decoding, where the future is absent for every row, and is the deployment metric. The masked Direct row is the dense-transfer decoder-Direct reference with the causal mask intact; unmasking it (unmasked Direct) keeps the CE level but collapses AR generation, and the mirror-Direct direction recovers both. Removing the mask alone is harmless only for dot (and slightly beneficial on both metrics). CE/PPL lower is better; ROUGE-L/BLEU-4 higher is better.

Arm (mask-pretrained Q/K)	Full-context CE ↓	PPL ↓	ROUGE-L ↑	AR BLEU-4 ↑
official pretrained dot (masked)	1.8414	6.31	0.3430	0.2297
masked dot (self-distillation)	1.8412	6.30	0.3439	0.2329
unmasked dot	1.8168	6.15	0.3475	0.2346
masked Direct (dense transfer)	2.409	11.12	0.3109	0.1897
unmasked Direct	2.3790	10.79	0.1697	0.0687
unmasked mirror-Direct	2.0637	7.88	0.3016	0.1732

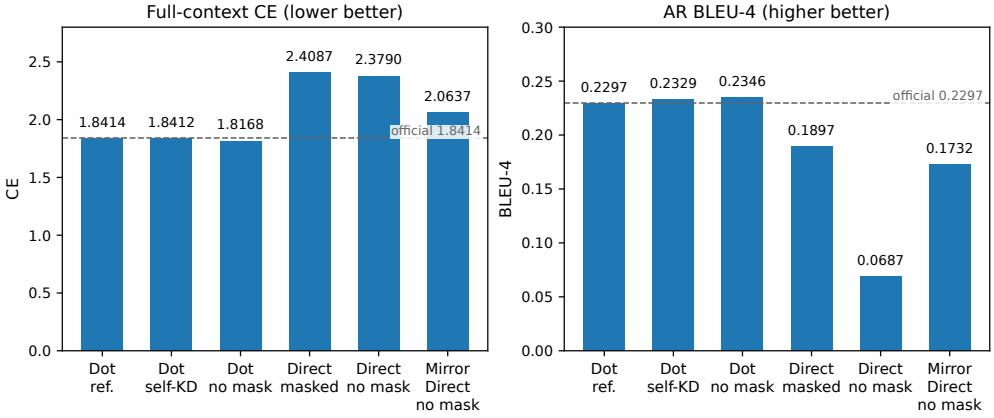


Figure P.1: Unmasked-attention ablation: eval CE (left) and AR BLEU-4 (right) for the official masked reference and the three mask-pretrained-Q/K arms of Table P.1. Dashed lines are the official reference.

Ablation conclusions (mirror method). Within this single-seed, 6,000-step study: (i) removing the mask leaves dot's AR behavior nearly unchanged while lowering its full-context CE; (ii) the same change for Direct is associated with a large AR BLEU drop; (iii) adding the mirror direction gives the best joint full-context-CE/AR-BLEU outcome among the tested unmasked Direct variants (eval CE 2.0637, AR BLEU 0.1732); and (iv) a residual gap to the masked reference remains. The supported claim is therefore protocol-specific: direction-flipped future scores improved the tested unmasked Direct baseline, without establishing that directionality is generally necessary or identifying the causal mechanism.

Q Random Q/K Initialization (rQK): Warm-Start Dependence under One Budget

Question and protocol. Every unmasked arm inherits the mask-pretrained Q/K, i.e. projections trained under the causal mask. A plausible failure mode is that the dense transfer sinks into the causal local minimum and never learns to use the future — which would mean the mask-pretrained init itself, not the operator, limits the no-mask results. The rQK choice ablation tests this by re-initializing the 16 decoder Q/K projections (kaiming, nn.Linear default) before the same 6,000-step DK; the column "rQK" in Table Q.1 marks which runs use the random init. Identity evidence motivates the operator split: on dot weights the unmasked identity costs 3.34 CE, on Direct weights about 10 CE regardless of the mirror direction — dot's bilinear scores retain a usable similarity structure even from random projections, Direct's subtract–rectify scores do not.

Ablation step 1: rQK on the Direct family. Random initialization performs poorly for both Direct-family arms under this single-seed, 6,000-step budget: unmasked Direct with rQK reaches eval CE 5.5415 / AR BLEU 0.0016 and unmasked mirror-Direct with rQK 3.5223 / AR BLEU 0.0000, against 2.3790 / 0.0687 and 2.0637 / 0.1732 from the mask-pretrained Q/K. The selection trajectories (Figure Q.1) show the observed optimization gap: the random-init arms start at selection CE 5.6–6.1 and recover only to 3.5–5.0 in 6,000 steps, while the mask-pretrained arms start at 2.4–3.1 and reach 2.36–2.56. Thus, successful adaptation was observed only with mask-pretrained Q/K in these tested Direct arms under this budget; the experiment neither establishes a universal warm-start requirement nor identifies score geometry as the cause.

Ablation step 2: rQK on dot. For the dot operator the observed pattern differs: unmasked dot with rQK reaches the best eval CE of the whole matrix (1.6693, -0.17 below the official, PPL 5.31), while its AR generation drops (AR BLEU 0.1954 versus 0.2346). The rQK dot arm reaches selection CE 1.74 by step 2,000 (Figure Q.1) and plateaus at 1.66. Within this single run, the lower full-context CE and lower AR BLEU are consistent with greater reliance on future information than in the mask-pretrained initialization. This descriptive contrast does not establish that the causal initialization generally constrains future use or that the same tradeoff persists across seeds and budgets.

Additional insights. Three further descriptive observations hold in the tested matrix. (a) Mirror-Direct has lower full-context CE than plain Direct under both initializations ($3.52<5.54$ with rQK, $2.07<2.38$ without), but one seed does not establish an initialization-independent advantage. (b) Across these equal-budget cells, lower full-context CE is associated with lower AR BLEU, consistent with a pairing-cost tradeoff. (c) Both Direct-family rQK arms recover more slowly than the dot rQK arm; this pattern is consistent with greater optimization difficulty from random projections, not proof of an operator-specific bottleneck.

Table Q.1: Random Q/K choice ablation (single seed, 6,000 steps, development evidence). The "rQK" column is the ablation choice (rQK = randomly initialized Q/K projections): ✓ = the 16 decoder Q/K projections were re-initialized (kaiming) instead of loading the mask-pretrained checkpoint's Q/K. The official row is the masked reference. Full-context CE is future-present for the unmasked rows and future-absent for the masked rows (see Table P.1); AR BLEU-4 is the future-absent deployment metric. CE/PPL lower is better; ROUGE-L/AR BLEU-4 higher is better.

Arm	rQK	Full CE ↓	PPL ↓	ROUGE-L ↑	AR BLEU-4 ↑
official pretrained dot (masked)	—	1.8414	6.31	0.3430	0.2297
unmasked dot	×	1.8168	6.15	0.3475	0.2346
unmasked dot	✓	1.6693	5.31	0.3214	0.1954
unmasked Direct	×	2.3790	10.79	0.1697	0.0687
unmasked Direct	✓	5.5415	255.06	0.0093	0.0016
unmasked mirror-Direct	×	2.0637	7.88	0.3016	0.1732
unmasked mirror-Direct	✓	3.5223	33.86	0.0011	0.0000

Final conclusions (unmasked-attention study). Organizing this single-seed, 6,000-step study: (1) removing the mask changes dot and Direct differently under the causal_only anchor; (2) the

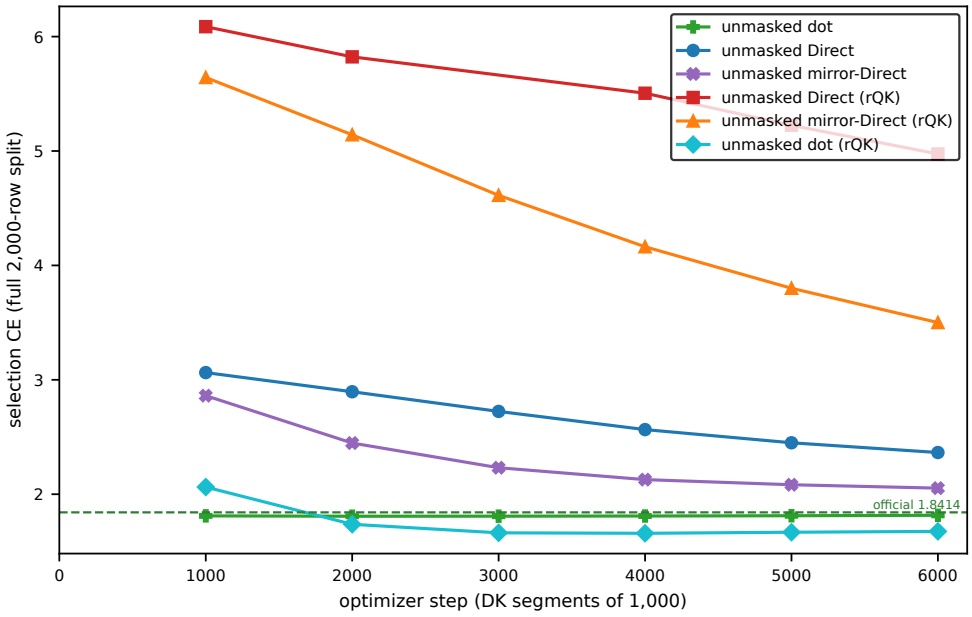


Figure Q.1: Random-Q/K recovery: per-segment selection CE (full 2,000-row split) against optimizer step for all six arms with segment evals (unmasked dot and unmasked mirror-Direct were re-run under the standard kernel-bf16 numerics to collect their trajectories; their final values agree with Table Q.1 to within 0.01 CE). The mask-pretrained arms start far lower; the random-Q/K arms recover slowly, except unmasked dot with rQK which reaches the lowest selection CE of the matrix.

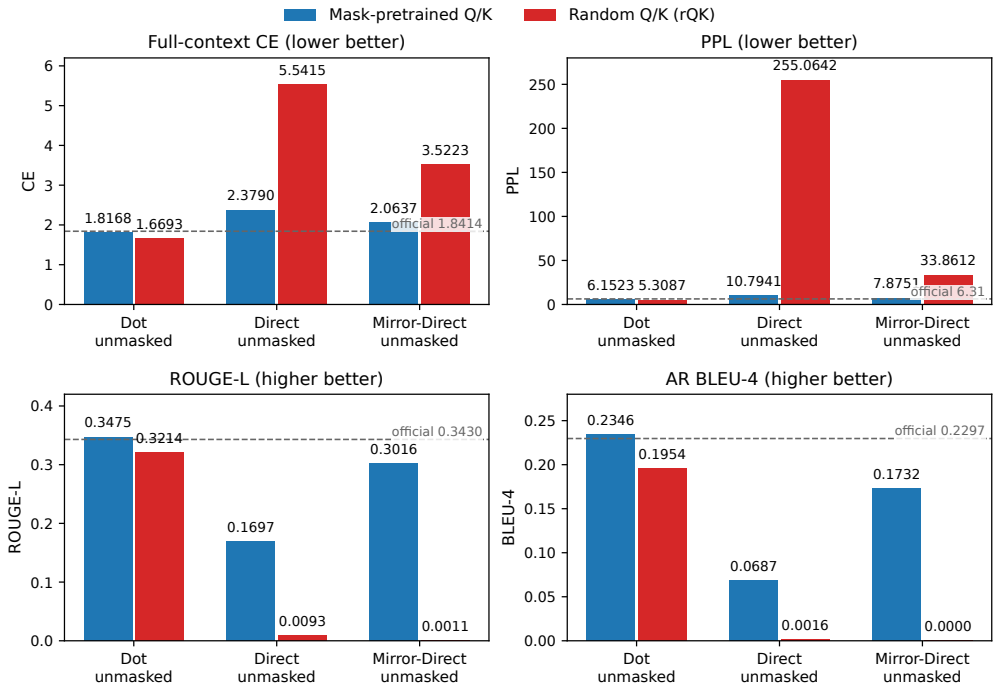


Figure Q.2: Final rQK results, four panels (eval CE, PPL, ROUGE-L, AR BLEU-4) with the seven matrix rows grouped by arm: blue = mask-pretrained Q/K, red = random Q/K (rQK). Dashed lines are the official masked reference.

mirror-Direct future direction gives the best joint outcome among the tested unmasked Direct variants, while retaining a gap of $+0.23$ CE and -0.057 AR BLEU to the masked reference; (3) successful adaptation of the tested Direct arms was observed only with mask-pretrained Q/K within this budget, without establishing a requirement for other methods or longer runs; and (4) full-context CE and autoregressive generation need not move together, so future-present modeling and future-absent generation must be reported jointly. These are development observations, not seed-robust superiority or causal-mechanism claims.

R Learnable Rectifier Parameters for Masked Direct Attention: an Eight-Cell Factorial Study

Motivation and protocol. Table P.1 shows the masked Direct gap (2.4087 versus the official dot 1.8414 under the fp32-fallback numerics). The zero-plateau diagnosis motivates three optional scalar families: a threshold θ_h, applied through paired shifts $q \leftarrow q - \theta_h/2$ and $k \leftarrow k + \theta_h/2$; a learned log-scale s_h; and a lower clip bound c_h. Each enabled family adds one scalar per head in each decoder layer, so the parameter count depends on the active families. The recovered source defines the current implementation; the historical factorial reports do not bind its hash and therefore do not attest that exact runtime source. Here s is not restricted to integer bit shifts. The threshold and clipping bound suggest add/compare primitives, but their implementation cost is not zero. A noninteger s generally requires approximating the fixed scale for a multiplier-free implementation; the approximation and its numerical effect are not evaluated in this study. The factorial uses the same masked 6,000-step DK protocol as Table P.1, with the 16 decoder Q/K tensors and selected new scalars trainable. All cells share the kernel-bf16 numerics; these are distinct from the W-bit/$2W$-bit integer contract in Appendix O.

Numerical base. The kernel-bf16 baseline (masked Direct, 6,000 steps, no rectifier parameters) reaches 2.6552 CE / 0.1845 AR BLEU-4. This differs from the fp32-fallback masked Direct of Table P.1 (2.4087 / 0.1897) by $+0.25$ CE. Because the intended contrast changes the score numerics, the result is consistent with sensitivity to bf16 rounding of fine $q - k$ differences; this single-seed comparison does not establish a general precision penalty or isolate every source of run-to-run variation. The factorial therefore compares within its own numerics, with the fp32-fallback row kept as a reference in Table R.1 and Figure R.2 (dashed line).

Results. Table R.1 gives the full factorial (rows grouped by the θ/s/c choice, columns mark which parameters are active). Figure R.2 shows the same eight cells for the full-context CE and the AR BLEU-4.

Factor analysis. Figure R.1 shows the per-segment selection CE for all eight cells: the s-containing arms (red/orange/purple families) separate from the s-free arms (grey/blue) within the first 2,000 steps and stay below the fp32-fallback reference line for the rest of the budget. Three descriptive patterns stand out within this single-seed factorial. (i) Every cell containing s improves both metrics (-0.25 CE, $+0.017$ AR BLEU-4 versus the baseline), while the measured s-free cells remain near the baseline. (ii) θ alone changes little (2.6546 versus 2.6552), whereas θ+s (2.3702) improves on s alone (2.4034) by a further -0.033 CE and $+0.002$ AR BLEU-4; this is suggestive of an interaction but is not a replicated factorial estimate. (iii) c, initialized at 16 and inactive for the observed score range $[-13, -1]$, receives no gradient while inactive and stays at 16; the c-containing cells therefore closely track their c-free counterparts. The lowest measured CE is obtained by θ+s: 2.3702 CE / 10.70 PPL / 0.3240 ROUGE-L / 0.2034 AR BLEU-4.

Learned parameter values and hardware mapping. The learned log-scales concentrate at $s \approx -0.23$ (range $-0.07 \ldots -0.24$ across layers and heads), i.e. $2^{-s} \approx 1.17$. For head h, the recovered implementation casts $\delta_h = \theta_h/2$ to the activation dtype and applies $q'_{hir} = q_{hir} - \delta_h$ and $k'_{hjr} = k_{hjr} + \delta_h$. It then computes

$$R_{hij} = -\sum_{r=1}^{d_k} \max(0, q'_{hir} - k'_{hjr}), \qquad U_{hij} = d_k^{-1/2} R_{hij},$$

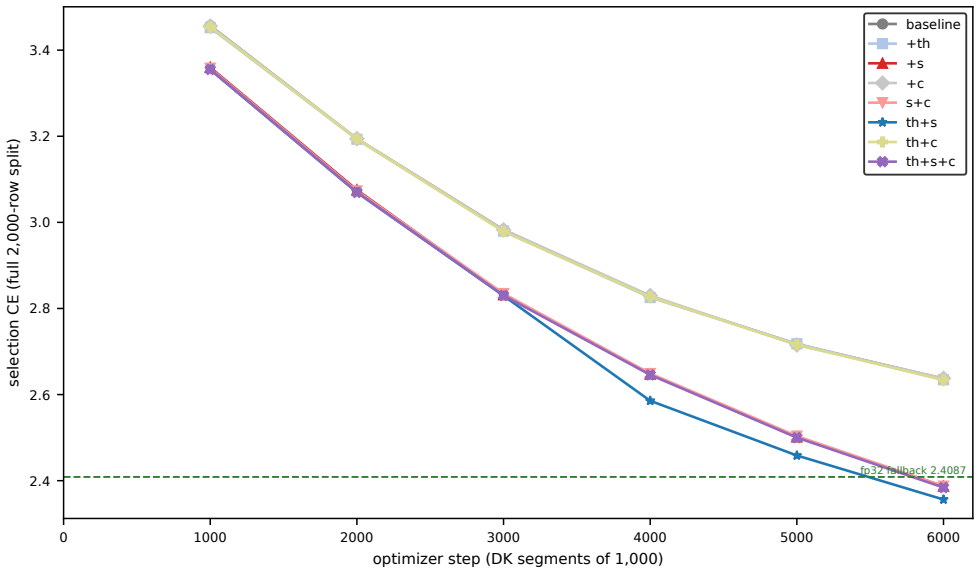


Figure R.1: Learnable-rectifier factorial: per-segment selection CE (full 2,000-row split) over the 6,000-step DK budget for the eight masked-Direct cells of Table R.1. The green dashed line is the fp32-fallback masked Direct reference (2.4087). The s-containing arms separate from the s-free arms within the first 2,000 steps; θ+s is the lowest trajectory.

followed, when enabled, by $V_{hij} = 2^{-s_h} U_{hij}$ and then $S_{hij} = \max(V_{hij}, -c_h)$; an inactive step is omitted. Masking follows these operations and precedes softmax. In exact arithmetic before clipping, $-\max(0, x - \theta_h) = -\max(\theta_h, x) + \theta_h$; after summation and scaling, the difference is the key-independent offset $\sqrt{d_k}\theta_h 2^{-s_h}$. This does not establish equivalence after activation-dtype rounding or the subsequent clamp.

An integer-exponent scale could be implemented by a shift; the observed fractional exponents instead require a specified finite-precision implementation. A shift/add approximation is a possible design, but its area, rounding error, and effect on model quality have not been measured. The causal-mask DK result (2.3702 / 0.2034) is the best masked-Direct configuration measured in this study; it is not an accuracy result for a synthesized implementation of the learned scales.

Caveats. Single seed, development split, 6,000 steps; the fp32-fallback vs. kernel-bf16 difference ($+0.25$ CE on the baseline) is a setup-specific precision-sensitivity result, not a general comparison with dot-product accumulation. The θ+s result stays above the official masked dot (1.8414), so no superiority claim is made; within the tested factorial, s is the factor most consistently associated with improvement.

Table R.1: Masked-Direct learnable-rectifier factorial (6,000 steps, single seed, development evidence; all cells share the kernel-bf16 numerics). θ is the rectifier threshold (per-head additive shift), s the learned log-scale (per-head 2^{-s} factor), c the lower clip bound (per-head scalar, STE-trained, initialized inactive). Full-context CE is future-absent here (masked rows); AR BLEU-4 is the greedy left-to-right generation metric. The dashed references in Figure R.2 are the fp32-fallback masked Direct (2.4087 / 0.1897) of Table P.1.

Arm	θ	s	c	Full CE ↓	PPL ↓	ROUGE-L ↑	AR BLEU-4 ↑
baseline	×	×	×	2.6552	14.23	0.3074	0.1845
$+\theta$	✓	×	×	2.6546	14.22	0.3070	0.1836
$+s$	×	✓	×	2.4034	11.06	0.3206	0.2019
$+c$	×	×	✓	2.6547	14.22	0.3079	0.1847
s+c	×	✓	✓	2.4034	11.06	0.3194	0.2019
θ+s	✓	✓	×	2.3702	10.70	0.3240	0.2034
θ+c	✓	×	✓	2.6538	14.21	0.3076	0.1833
θ+s+c	✓	✓	✓	2.3992	11.01	0.3216	0.2018

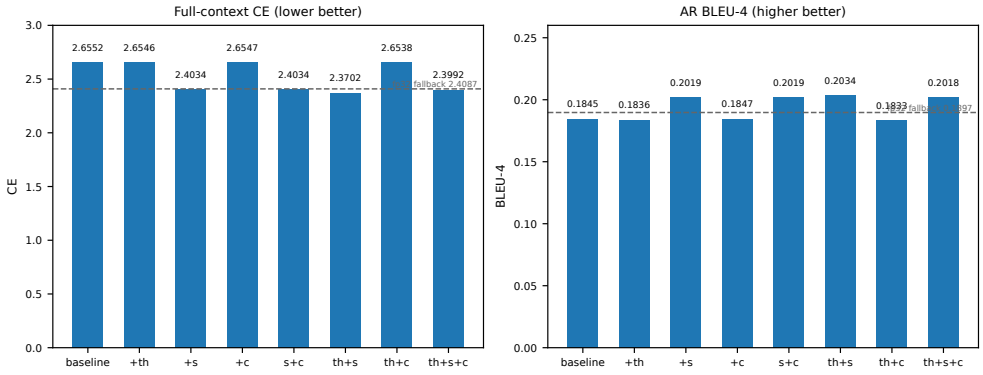


Figure R.2: Learnable-rectifier factorial: full-context CE (left) and AR BLEU-4 (right) for the eight masked-Direct cells of Table R.1. Dashed lines mark the fp32-fallback masked Direct reference (2.4087 / 0.1897). In this single-seed matrix, only cells containing the log-scale s improve both metrics; the θ+s cell has the lowest measured CE.

S Direct Attention in a Modernized Diffusion Transformer for Class-Conditional Image Generation

Motivation and model. The standard class-conditional DiT uses bidirectional self-attention (Peebles & Xie, 2023), making it a useful testbed for Direct without a decoder-style causal mask. We build a self-contained *Modernized DiT-S/2* (latent $[B, 4, 32, 32]$, patch 2, 256 tokens, 12 layers, 6 heads, hidden 384, 33M parameters, class-conditional on the 1,000 ImageNet classes) with QK-Norm, 2D RoPE, Pre-Norm, and bias-free linear layers. The attention operator is switchable between the standard dot score and the Direct score $-\sum_d \max(0, q_d - k_d) \cdot D^{-1/2}$. The archived program defaults to a separate DiT-specific bf16 CUDA implementation and provides a chunked native-PyTorch fallback; the historical run record does not preserve the fallback environment flag, so the reported DiT curves are used only as operator-level feasibility evidence, not as kernel-performance evidence. A post-hoc, source-hashed audit of the archived DiT kernel compared 7,806 forward values and 38,912 query/key gradient values across three D64 cases, including ties, against an fp32-formula oracle rounded to bf16; no value differed by more than one bf16 ulp. This DiT-specific audit does not rehabilitate the failed Stage 3 backend above.

Protocol. Two arms (dot, direct) train on ImageNet-1K latents (SD VAE, frozen, online-encoded into a shared fixed-order disk cache so both arms consume the identical augmented stream): 5,000 iterations, batch 64 (320k samples, 0.25 epoch), lr 10^{-4} with 500-step warmup and cosine decay,

DDPM simplified loss (Ho et al., 2020), bf16 autocast with fp32 parameters, fixed seed 42 with per-step deterministic noise/timestep RNG. This is a fixed compute budget, not a convergence budget; single-seed development evidence only.

Results. Table S.1 reports the simplified-loss trajectory; Figure S.1 shows the full curves. The direct arm is learnable under this protocol: its final loss (0.1926) is within 2.3% of the dot arm (0.1971, ratio 0.977), both starting from 0.9985. Figure S.2 shows DDIM-50 samples (Song et al., 2021) (eta 0) at the end of the budget: color-coherent, class-differentiated early-stage outputs for both operators, consistent with the 0.25-epoch budget (recognizable objects require orders of magnitude more steps for either operator).

Judgement. Under this short protocol, both arms train without divergence and finish at similar fixed-set losses. The evidence is limited to one seed, 0.25 epoch, and a training loss; it includes no FID, EMA, or guidance study. It therefore demonstrates numerical learnability in this small DiT setup, not comparable generation quality or parity at convergence.

Table S.1: Modernized DiT-S/2 on ImageNet-1K latents: DDPM simplified loss (mean squared prediction error), 5,000 iterations, batch 64, single seed. The Direct/dot final-loss ratio is 0.977; judgement: Direct learnable (working).

Arm	init loss	final loss (5,000 steps)
dot	0.9985	0.1971
direct	0.9985	0.1926

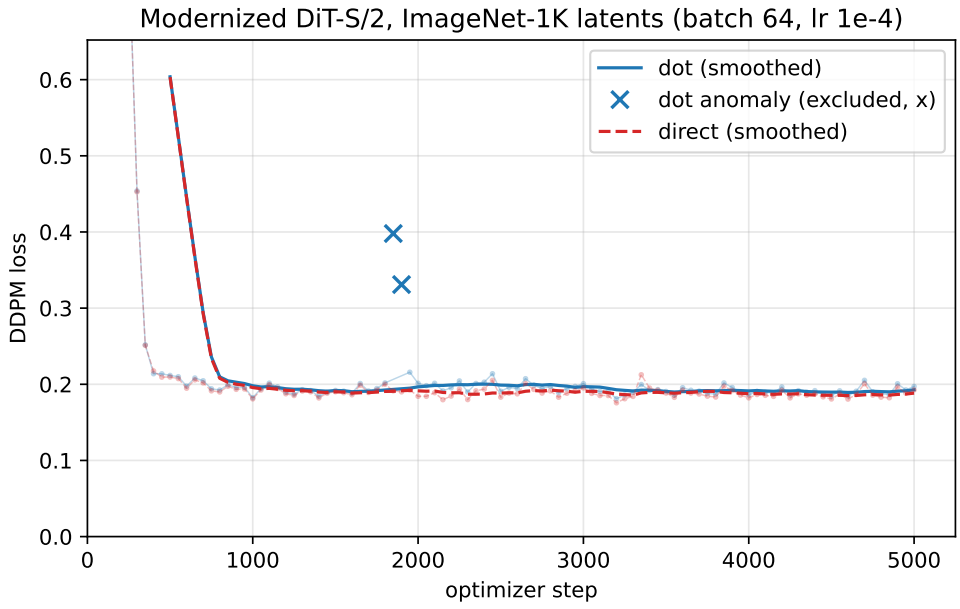


Figure S.1: Training curves (linear-scale loss, same schedule and identical augmented data stream) for the dot and direct arms of Table S.1: both start near 0.9985 and descend to within a few percent of each other at the 5,000-step budget, showing stable end-to-end optimization for both operators. Two dot-arm points (steps 1,850 and 1,900) are a known experimental anomaly: they are marked with $\times$ and excluded from the plotted lines.

Cross-regime assessment. The short from-scratch study provides a bidirectional-generation training diagnostic alongside the causal-masked and checkpoint-transfer studies. Its fixed-set loss proximity should not be extrapolated to sample quality or converged performance.

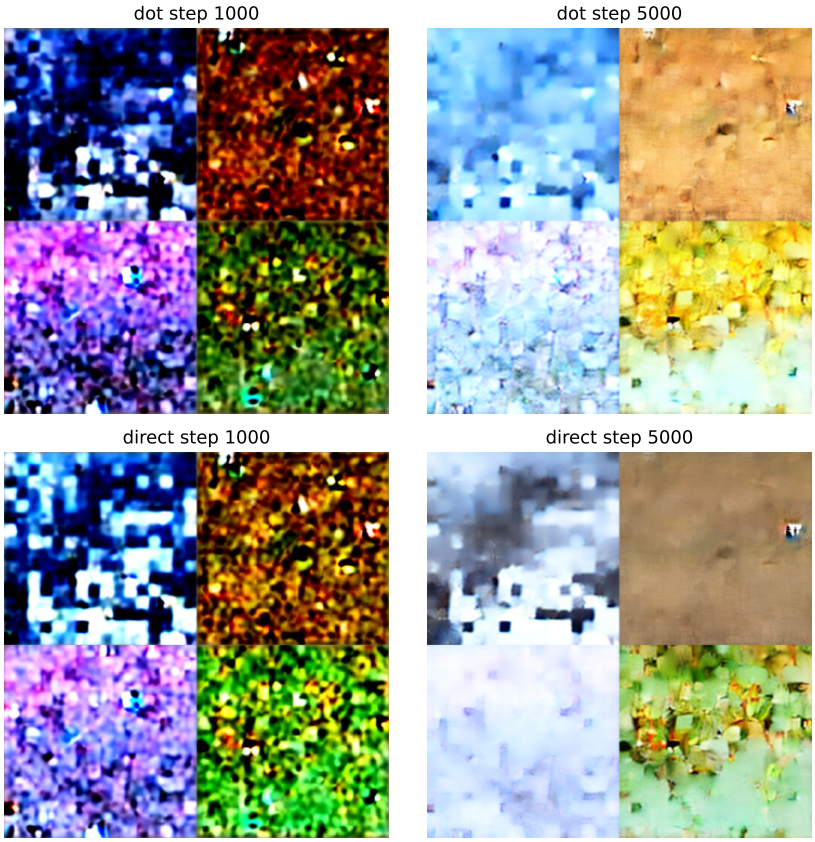


Figure S.2: DDIM-50 samples (eta 0, fixed classes/noise) at the 5,000-step checkpoints: color coherence and class differentiation are visible for both operators at the 0.25-epoch budget; recognizable object structure is not expected at this budget for either arm.

T EXCLUDED PRETRAINED-DIT TRANSFER ATTEMPT

Audit disposition. This attempted pretrained DiT-XL/2 transfer is excluded from the paper's evidence. During final source audit, both the retained training and evaluation programs were found to construct the noised latent as $x_t = \bar{\alpha}_t x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$, whereas the stated DDPM schedule requires $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$. Consequently, the attempted transfer's checkpoints, endpoint losses, per-timestep diagnostics, and generated samples do not evaluate the claimed protocol and are not used to support any conclusion. No corrected rerun is included. The separate from-scratch DiT experiment in Appendix S uses the correct square-root coefficient and is unaffected. We retain this short disclosure so that the omitted experiment is not mistaken for selectively withheld negative evidence.

U LIMITATIONS AND SCOPE

Primary ImageNet result. The main accuracy result is the completed, fixed-budget Q/K conversion in Appendix A, with endpoint the Phase 3 terminal checkpoint at local optimizer step 5000, initialized from the selected Phase 2 source at global optimizer step 35000. It retains a residual gap to the frozen Dot reference. The reference is an internal checkpoint baseline, not an external state-of-the-art benchmark. The single adaptive trajectory and reused validation split limit inference to this reported endpoint, as detailed in Appendix B; the paired image-level interval does not quantify training-seed uncertainty or establish parity or non-inferiority.

Conversion versus symmetric training. The source checkpoint was pretrained on ImageNet-21K and fully fine-tuned on ImageNet-1K using Dot. The conversion then changes only Q/K weight rows on ImageNet-1K, while all remaining parameters and the Dot teacher stay frozen. The complete objective, parameter accounting, and phase-specific schedules appear in Appendix A. This design measures constrained checkpoint conversion, not operator capacity under equal training compute; it also provides no training-speed comparison of dense supervision, ordinary fine-tuning, and output-only distillation.

Legacy mechanism and explainability results. The older bias-preserving checkpoint pair, not the primary conversion endpoint, underlies the retrieval and explanation studies. For that pair, shared geometry and within-model component and patch-removal tests coexist with failed threshold-ranking, connectivity, and localization gates. Thus the tested functional-reparameterization hypothesis is rejected; the failed low-dimensional path does not rule out every nonlinear parameter-space path. The interventions hold V, projection, MLP, and classifier weights fixed and test immediate inference response rather than training compensation. Explainability covers 17 dense spatial methods on a jointly-correct 1,000-image set, excluding LIME, SHAP, TCAV, and concept-based methods.

Implementation scope. The hardware result is a Sky130 HD mapped-area comparison of the score stage, using the same frozen-Dot traces under generic preprocessing for both units. The scripts specify TT, 25 °C, and 1.8 V, although exact historical library/tool identities remain incomplete. SRAM-inclusive totals are a separate analytical ledger. Neither is a silicon measurement or an end-to-end timing or energy result, and the comparison does not establish a speed advantage over GPU matrix multiplication. Value aggregation, projections, and MLPs still use multiplication. Historical timing and power records remain excluded diagnostics (Appendix O).

Coverage and unresolved diagnostics. Classification evidence is limited to ViT-Tiny and ViT-Small, with no independent multi-seed repeats. MiniMind-V and NLLB-200 likewise provide single-seed development results rather than confirmatory comparisons. The exploratory Tiny search's historical 7.33% anomaly remains unresolved (Appendix L). The bias-deleted KD checkpoint and failed Stage 1 gate retain their historical status (Appendix K); the invalid pretrained-DiT transfer is excluded (Appendix T). None is pooled with the primary conversion result.