WHAT DOES ATTENTION TRANSFER TRANSFER? ATTENTION STRUCTURE AND ROBUSTNESS IN VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision transformers (ViTs) trained to copy a pretrained teacher's attention maps recover most of fine-tuning's in-distribution accuracy yet fall measurably short of it under distribution shift, as recent work has shown. What the copy delivers has never been measured directly in the attention structure and tied to robustness. We build that instrumentation for ViT-S students of a self-supervised teacher on ImageNet-100, and report three findings that triangulate one conclusion. First, the transfer is essentially perfect and permanently so: the distilled student's attention ends up roughly two orders of magnitude closer to the teacher's than fine-tuning does, and does not drift with additional training. Second, the gap is real at 14$\times$ fewer parameters and 10$\times$ less data than previously studied, but it has a time axis. It tracks training maturity, and completing the schedules that the stopping rule interrupted closes it below our pre-registered threshold in two of three seeds, with comparisons at equal accuracy giving the same result. The endpoint gap at this scale is substantially a training-maturity artifact: robustness matures later than accuracy, and stopping rules tuned to accuracy undersample it. Third, forcing cross-row redundancy down by half the structural separation between the distilled and fine-tuned conditions produces no detectable robustness response under two registered ways of matching accuracy. Verified transfer, a gap that closes while the structure never moves, and a null under direct intervention are together consistent with the deficit residing in features, not in the visible attention structure. This is elimination plus intervention, and its scope is the regime we measured. In this regime, attention overlays show where a model looks, not what it knows.

1 INTRODUCTION

Attention maps are among computer vision's most trusted windows into a transformer. They are overlaid on images to show a model "attending to the object" (Wu et al., 2024), scored for quality (Zhou et al., 2022; Darcet et al., 2024), and distilled from teacher to student (Zagoruyko & Komodakis, 2017) on the premise that they carry what the teacher knows. Yet a student trained to copy a pretrained teacher's attention maps recovers most of fine-tuning's clean accuracy while remaining measurably less robust to distribution shift (Li et al., 2024), a gap conjectured but never tested to lie in the features. If the attention arrives intact, why doesn't the robustness?

This paper builds the instrumentation to answer that question, and the answer reframes it. Attention transfer copies what an attention map shows, the routing that decides which tokens read from which, while the student learns its own features from scratch. We train scratch, fine-tuned, and attention-distilled ViT-S students (the latter two inheriting from a single self-supervised teacher), and measure what attention transfer actually transfers, for the first time directly in the attention structure and tied to robustness. Three findings compose the answer.

First, **the copy is essentially perfect, and permanently so.** In per-layer, per-head KL divergence on held-out images, the distilled student's attention sits on average 97$\times$ closer to its teacher's than fine-tuning ends up from the same teacher, though fine-tuning began from the teacher's own weights. The structure holds across stopping epochs spanning half the schedule. Whatever explains the gap, it is not a failed copy.

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.

Second, **the gap is real but has a time axis.** It reproduces at 14× fewer parameters and 10× less data than Li et al. (2024), under the pre-registered rule and horizons. Across all fifteen distill-family runs, endpoint effective robustness rises with stopping epoch at $r = 0.978$; the gap shrinks as distillation approaches schedule completion while the attention structure never moves. Full-schedule reruns of all three distilled seeds bring the gap below our pre-registered threshold of 0.01 in two of three, and comparisons at equal accuracy give the same result. The endpoint robustness gap at this scale is substantially a training-maturity artifact.

Third, **structure responds to force; robustness does not follow.** An attention diversification penalty (Guo et al., 2023) at increasing dose moves the transferred structure monotonically through half the structural distance between conditions, with sharpness, fidelity, and completed-schedule accuracy held. Turning this causal dial produces no detectable robustness response under either of two registered ways of matching accuracy. The deficit is consistent with residing in features, not in the visible attention structure.

Contributions. (i) The first direct measurement of what attention-map distillation does to attention structure, tied to robustness: transplant fidelity, maturity-invariance, and the two-senses dissociation of "overfocusing." (ii) Reproduction of the gap under pre-registration, with its maturity structure and within-seed closure at completed schedules, revising how fixed-budget robustness comparisons should be read. (iii) A causal dose-response on transferred attention structure, with a robustness null at matched skill under a displacement of half the condition separation. (iv) A measurement methodology of matched-accuracy coordinates, stopping-variance quantification, and continuation calibration, reusable wherever early stopping meets no-peak curves.

2 Related Work

Attention transfer and what distillation carries. Transferring a teacher's attention is one of the oldest ideas in knowledge distillation (Zagoruyko & Komodakis, 2017); its modern ViT form trains a randomly-initialized student to match a pretrained teacher's attention maps. Li et al. (2024) showed that such a student recovers most of fine-tuning's in-distribution accuracy while falling short on distribution shift; they measured that gap at fixed budgets, with no stopping rule and no convergence evidence, and conjectured but did not test it to be feature-borne. We build the instrumentation such a test requires, and the budget condition turns out to matter (Section 4.2). Li et al.'s own appendix makes a first gesture in this direction (a layer-wise JSD between attention maps, with an acknowledged identifiability caveat); we replace it with a purpose-built multi-axis panel tracked across training, seeds, and doses, tied to robustness. A parallel line asks what distillation actually transmits (Stanton et al., 2021; Ojha et al., 2023); concurrent work by Qin et al. (2026) uses per-layer student–teacher attention divergence as a diagnostic for when attention transfer fails across architectures. We differ in object and method: we measure the *structure* of what arrives (concentration, redundancy, stability), show it is transplanted essentially intact and held across training, and then intervene on it causally. Where our results overlap with Qin et al.'s diagnosis, they agree: faithful attention does not guarantee transferred competence. Distillation under shift has been benchmarked (Zhang et al., 2025) without studying training-duration effects, where our result lives.

Attention pathologies in ViTs. Two distinct notions of "bad attention" circulate, usually unseparated: per-row collapse, attention entropy imploding as tokens fixate (Zhai et al., 2023), and cross-row redundancy, every query attending to the same few tokens, identified by Guo et al. (2023), whose Attention Diversification Loss we adopt as our intervention. No prior work separates them; our measurements show they dissociate across training conditions and can be manipulated independently, recasting "overfocusing" as two-dimensional. Unlike diversity interventions aimed at generalization (Nicolicioiu et al., 2023), our use of ADL is diagnostic, not performance-seeking.

Attention as evidence. NLP litigated whether attention weights explain predictions (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019); vision has since built faithfulness metrics for attention-derived saliency (Wu et al., 2024). That thread concerns the prediction-explanation axis. Ours is different: whether attention structure carries *robustness*, probed with a transplant design in which the "good-looking" attention is held fixed by construction while the property it supposedly certifies fails to arrive. The reading we counter is live practice: attention-map properties are credited for

robustness (Zhou et al., 2022) and artifact-free maps engineered as desirable in themselves (Darcet et al., 2024).

Robustness dynamics during training. Effective robustness (Taori et al., 2020; Miller et al., 2021) is known to evolve. Andreassen et al. (2022) showed pretrained models' effective robustness erodes over fine-tuning; Lin & Heckel (2024) describe late-training distributional overfitting; adversarial robustness has its own overfitting dynamics (Rice et al., 2020), and a rhyming decay of teacher-alignment value with training maturity appears in diffusion training (Wang et al., 2025). In our distillation regime the ratio-form quantity we report rises with maturity ($r = 0.978$); we do not claim a sign reversal against Andreassen et al., whose quantity is the fitted-baseline residual, under which our gradient itself reads negative (Appendix G); we report both directions. The lesson is that fixed-budget robustness comparisons are regime-dependent measurements of a moving quantity, and can invert.

Run-to-run variance and underspecification. Our stopping-lottery finding extends known variance from stopping and seeds (Dodge et al., 2020) and pipeline underspecification (D'Amour et al., 2022); our contribution is the composition on no-peak curves, where structure is stop-invariant, endpoint robustness is not, and the matched-accuracy coordinate follows.

3 Setup

Registered hypotheses. The registration fixed three directional predictions: that distilled attention would be *more* overfocused than weight-transferred attention; that the layer-wise overfocusing signature would predict the ID-to-OOD gap; and that diversifying the distilled structure would close the gap. They are adjudicated against their registrations in Section 4.1, Appendix A, and Section 4.4, respectively.

Conditions. We compare three ways of training a ViT-S/16 classifier (Dosovitskiy et al., 2021; Touvron et al., 2021) on ImageNet-100 (pipeline overview: Figure 4, Appendix I): training from random initialization (*scratch*); initializing every weight from a self-supervised teacher and fine-tuning (*weight transfer*, the "fine-tune" condition of Li et al. 2024); and training from random initialization while distilling only the teacher's attention maps (*attention transfer*, "distill"). The distillation term is a per-layer, per-head KL divergence between student and teacher attention distributions, added to cross-entropy at $\lambda_{\text{kl}} = 36$: not a tuned hyperparameter but the arithmetic equivalent of the official per-layer strength of Li et al. (2024) in our mean-over-layers formulation, verified to reproduce the official implementation's gradients to float precision (Appendix B). The causal intervention of Section 4.4 adds the Attention Diversification Loss (ADL) of Guo et al. (2023) at doses $\lambda_{\text{adl}} \in \{0.1, 0.3, 1.0, 3.0\}$, applied per-head to the same attention quantity our diagnostics measure (Guo et al.'s Eq. 1–2). Three seeds per condition throughout.

Teacher. The teacher is a ViT-S masked autoencoder (He et al., 2022) pretrained from scratch on ImageNet-100 with probe-based early stopping (Appendix I); both transfer conditions inherit this single checkpoint, so teacher quality cannot differentiate them (Section 6).

Training protocol. All conditions train under one pre-registered early-stopping rule, inside per-condition cosine horizons: validation top-1 every 5 epochs, min-delta 0.1pp, patience 5 evaluations (25 epochs of demonstrated flatness). The random-init family runs at 300 epochs (the standard DeiT-style schedule, fixed before any run); fine-tuning at 200, its initial 100-epoch schedule doubled once under a pre-declared pilot-calibration clause. Runs either stop early under the rule or complete their horizon; the latter are reported at their best checkpoint with the cap-hit flagged and terminal flatness disclosed (Appendix C). Convergence claims never rest on the rule's verdicts but on completed or completion-bounded anneals, shown terminal flatness, seed replication, and measured interruption cost (Section 4.2). Checkpoints are retained every 5 epochs, full precision, no pruning: every accuracy level any run passed through remains a loadable model, enabling the matched-accuracy analyses. Full configuration, hardware, and compute accounting are in Appendix H.

Evaluation. ID accuracy is top-1 on the held-out ImageNet-100 validation set. OOD evaluation uses the natural-shift suite ImageNet-A/R/Sketch/V2 (Hendrycks et al., 2021b;a; Wang et al., 2019;

Recht et al., 2019), label-mapped to the covered classes (15/19/100/100 of 100; suite examples: Figure 5, Appendix I). We report *effective robustness*: OOD accuracy divided by ID accuracy on the covered classes; of the skill a model has, the fraction that survives distribution shift. This ratio differs from the residual-above-baseline definition of Taori et al. (2020) and Miller et al. (2021); we flag the collision explicitly. Our primary comparisons are checkpoint-matched or near-equal in ID, where the two definitions order models identically; where ID differs (the endpoint coordinate and the maturity correlation) they can disagree, and Appendix G reports the residual-form recomputation. Uncertainty on endpoint quantities comes from a paired image-level bootstrap ($B = 10{,}000$, percentile 95% intervals); matched-coordinate quantities on trajectory checkpoints instead carry the single-pair noise scale ($\pm$0.5–1pp) and disclosed seed spreads (Appendix I). A frozen, self-generated ImageNet-C (Hendrycks & Dietterich, 2019) provides a secondary synthetic-shift probe (Appendix J). Attention structure is measured on the validation set with three quantities: per-row entropy and top-k mass (concentration); cross-row redundancy, Guo et al.'s Eq. 1–2 per head (whether different tokens stare at the same places); and clean-corrupt cosines of attention maps and Q/K projections (stability under corruption). All selection decisions use ID validation only; OOD suites are touched exactly once per evaluated checkpoint.

Comparisons. We use two co-primary coordinates, both fixed in the analysis code before any training run. The *endpoint* coordinate takes each method's rule-governed best, closest to the literature's practice, and is always reported with the qualifier "under the pre-registered rule and horizons." The *matched-accuracy* coordinate compares models at common ID accuracy using trajectory checkpoints. A calibration layer of full-schedule reruns (*continuations*: same seed and schedule from epoch 0, stopping disabled, replay fidelity checked against the canonical prefix) prices the rule's interruptions. They are labeled artifacts that never replace canonical numbers: first-class evidence, second-class authority.

4 Results

4.1 The transfer is faithful

Attention transfer transfers attention. The direct measurement is eval-time divergence on held-out validation images (Table 11, Appendix K): the distilled student's attention sits at a mean per-layer, per-head KL of 0.017 nats from its teacher's (0.012–0.022 across seeds, map cosine 0.99), against 1.65 nats for fine-tuning and 2.15 for scratch. The student trained to copy the teacher's gaze ends up 97$\times$ closer to it than the student that *started from the teacher's own weights*: the transplant held while the inheritance wandered. Per-layer KL never exceeds 0.042 nats in any layer of any seed (Figure 6, Appendix K), so the mean is not hiding a broken layer. Every other structural measure agrees (Table 12, Appendix K): mean redundancy lands within 0.002 of the teacher's value, about 1/25 of fine-tuning's drift (worst seed +0.0026), entropy and top-k mass are pinned, and corruption-stability *exceeds* every other student's.

This faithfulness is the study's fixed point: it replicates across seeds and is invariant to training duration, with students whose stopping epochs span 145–290 and whose accuracies span 5.6pp carrying the same structure to within measurement noise (redundancy $\pm$0.0008, entropy $\pm$0.006). Structure locks in early and permanently; everything else keeps maturing around it (Section 4.2). The explanation for the robustness gap cannot be a failure to copy: the copy is verified, stable, and, by the standards the literature uses to score attention quality, the best-looking attention in the family (overlays in Appendix K).

Two corollaries frame everything downstream. First, the two senses of "overfocusing" dissociate: transferred attention is the *most* concentrated per-row (entropy 3.61 vs. fine-tune's 4.14 and scratch's 4.34) and the *least* redundant across rows (0.468 vs. $\sim$0.51 and $\sim$0.58). We had registered the prediction that transferred attention would be more overfocused than weight-transferred attention. Per row it is; across rows it is the reverse. The prediction was not answerable as posed, which is the finding. Second, supervised training *increases* redundancy when left free: fine-tuning drifts from the teacher's 0.4665 to $\sim$0.51; scratch, never anchored, reaches $\sim$0.58 (Figure 8, Appendix K).

4.2 THE GAP IS REAL, AND IT HAS A TIME AXIS

The robustness gap reproduces. At rule-governed endpoints, attention-transfer students trail weight-transfer students in effective robustness at every seed: gaps of 0.0575 [0.0445, 0.0703], 0.0323 [0.0209, 0.0442], and 0.0181 [0.0073, 0.0289] against a pre-registered threshold of 0.01 (Table 3, Appendix F; paired bootstraps: every interval excludes zero and every point estimate clears the threshold, though s2's interval reaches below it). Endpoint gaps are reported here and throughout under the pre-registered rule and horizons. This reproduces the ordering of Li et al. (2024) with a 14× smaller student on a 10× smaller dataset: weight transfer above attention transfer above scratch, at every seed (endpoint table in Appendix F).

But the per-seed gaps are not three estimates of one number. They are three samples of a *curve*. The distilled stopping epochs ranged from 145 to 290 (rule variance characterized in Appendix E). Endpoint robustness tracks stopping epoch at $r = 0.978$ across all fifteen distill-family runs (3 dose-zero, 12 dosed; Figure 1A), using ratio-form effective robustness throughout (Appendix G reports the residual-form recomputation, under which this gradient's sign reverses; the study's two registered verdicts rest on matched- and near-equal-ID comparisons, where the definitions agree). The relation is not an artifact of pooling (restricted subsets give 0.973–0.977, including the subset with all horizon-capped runs dropped). Later-stopping runs are more robust. The endpoint gap shrinks monotonically as distillation approaches schedule completion, $0.057 \rightarrow 0.032 \rightarrow 0.018$, while the attention structure never moves (Section 4.1). Whatever matures is not the routing.

Three clocks. A densified pass over one seed's saved trajectory separates three timescales (Figure 1B). Structure locks first (redundancy at its final value by epoch 24, ± 0.002 through 299); accuracy next (93% of final by epoch 119); robustness last (highest probe at epoch 294 of 300). In the 55-epoch window bracketing the rule's cut, the same seed gained 3.5pp of effective robustness against 1.2pp of ID accuracy, and the two later seeds' full remainders are same-signed (+3.1 vs. +1.8pp, +0.7 vs. +0.3pp; Appendix D).

Completing the schedules. The maturity reading predicts a far smaller gap at *completed* schedules. We tested this within-seed: for each distilled seed, a full-schedule rerun extends the canonical run to its 300-epoch horizon. Replay was exact across 131 shared evaluations, so each rerun *is* the canonical run continued. The interruption cost is a dose-response in how hot the rule cut: the seed cut at 59% of peak learning rate recovered +6.0pp ID and +5.3pp effective robustness; at 20%, +1.8 and +3.1pp; at 1.3% (an in-design negative control, cut 10 epochs from the horizon), +0.3 and +0.7pp. Meanwhile the transferred structure stayed immobile through all three full schedules (redundancy 0.4673/0.4676/0.4673, still at the teacher's 0.4665).

At completed schedules the gap closes to below the pre-registered threshold in two of three seeds: 0.0044 [−0.0059, +0.0148], 0.0013 [−0.0096, +0.0122], and 0.0110 [+0.0003, +0.0217]. The first two are statistically indistinguishable from zero at image-level noise; the third sits at 1.1× the threshold and pairs the strongest fine-tuned seed with the weakest completed distilled one. The completed-schedule bands are adjacent: distill 0.4854–0.4901 vs. fine-tune 0.4913–0.4964, about 0.55pp between means. The endpoint and completed-schedule triples are never blended: one is what a practitioner's stopping rule sees, the other what the methods produce at equal, completed budgets. The conclusion this licenses: **the endpoint robustness gap at this scale is substantially a training-maturity artifact**. Under the pre-registered rule and horizons the gap is real and in line with Li et al.'s; at completed schedules it compresses below threshold in two of three seeds, with equal-accuracy comparisons agreeing (Section 4.3). For seed 0, 92% of the endpoint-coordinate gap ($0.0575 \rightarrow 0.0044$) was maturity, not method.

4.3 MATCHED-SKILL COMPARISONS

The endpoint coordinate compares models that reached different accuracies; the matched-accuracy coordinate removes that, comparing fine-tuned and distilled trajectory checkpoints of the same seed at common ID accuracy, at each seed's own matched level and at the top level both reach (Table 5, Appendix F); the distilled top rungs come from the labeled reruns.

At low matched levels the per-seed gaps are large *and of both signs*: +0.0237 at 0.8248 (s0), −0.0145 at 0.8618 (s1, the study's one loose match: its fine-tuned side sits 0.0054 below target, inflating

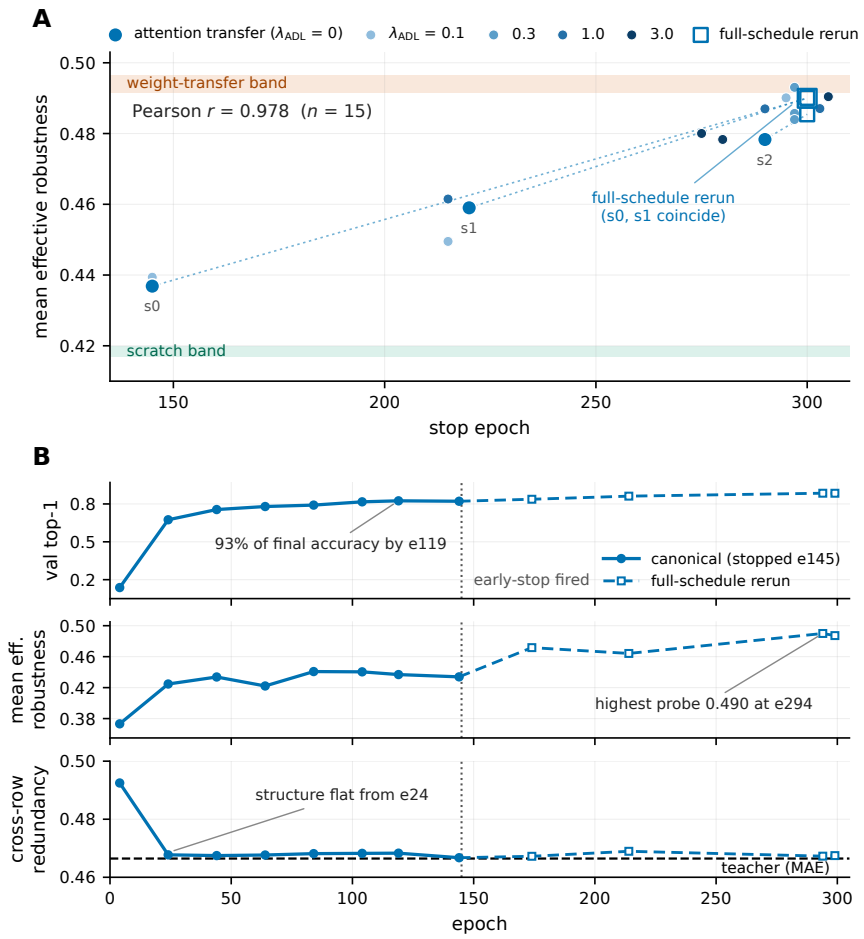

Figure 1: **A:** endpoint effective robustness vs. stop epoch, all 15 distill-family runs ($r = 0.978$; bands: min–max of weight-transfer and scratch seeds; open squares: full-schedule reruns at epoch 300; cap-stopped runs dodged $\leq$5 epochs for legibility). **B:** three clocks within seed 0 (canonical solid, rerun dashed): structure flat from epoch 24, accuracy at 93% of final by 119, robustness's highest probe at 294; dotted line = the rule's cut at 145.

this negative gap by roughly 0.6pp; Appendix F), +0.0049 at 0.8812 (s2). The spread is not dose or method but robustness-arrival heterogeneity in the baselines themselves (at the common level 0.8248 the three distilled seeds' effective robustness spans 4.3pp), so a single-level matched read is seed-noisy by construction, its registered mean of -0.8pp a candidate null we decline to summarize with one number.

Maturity resolves the heterogeneity. At the top matched rungs ($\sim$0.88, levels only completed schedules reach) the gaps are +0.0037, -0.0030, +0.0049: mean +0.002, all inside the $\pm$0.01 threshold band, still mixed in sign. (The one negative sign rests on a registered selection detail and reads +0.0016 under the alternative selection; all three stay inside the band either way.) Within-seed, the matched gap converges toward zero from both directions as level rises: s0 from +2.4pp to +0.4pp, s1 from -1.5pp to -0.3pp (Figure 9, Appendix K). Skill-matching at mature levels agrees with the completed-schedule endpoint read. One honesty bound: the fine-tuned top-rung checkpoints sit at $\sim$75–82% of their annealing schedule. In the one measurable case (fine-tuned seed 0's mid-anneal epoch-129 checkpoint, ID 0.8710, against the retired 100-epoch horizon's fully-annealed endpoint, 0.8718; Appendix C) the less-complete schedule scored *higher* by 2.0pp at matched ID, so this bias inflates rather than hides the gaps, making them conservative; but the ladder cannot resolve differences below $\lesssim$1pp, which is why the completed-schedule endpoint read, with both sides fully annealed, carries the primary weight (Appendix F).

4.4 THE CAUSAL TEST

Everything so far is observational. The intervention asks whether the structure *causes* any of the deficit. Our registered prediction was that it does: that diversifying the transferred structure would narrow the gap. Adding Guo et al.'s diversification penalty to the distillation objective at doses $\lambda_{\text{adl}} \in \{0.1, 0.3, 1.0, 3.0\}$ turns the one visible knob: cross-row redundancy (Figure 2).

The dial turns, and it is a scalpel. Paired same-seed displacement of cross-row redundancy is monotone in dose and accelerating: -0.0025, -0.0038, -0.0100, -0.0247 (dose means, n=3 each, per-seed values within 0.0011 of their dose mean; full table in Appendix F). The top dose moves redundancy by half the entire distill–fine-tune condition separation (0.0247 of 0.0445), near-deterministically across seeds ($\pm$0.0002). The instrument is one-directional: ADL diversifies, pushing redundancy *away* from fine-tuning's higher value and below the teacher's; the separation fraction is a magnitude yardstick, not a statement that the student moved toward fine-tuning's structure. To first order, nothing else moves: per-row entropy and top-k mass stay family-flat at every dose (maximum entropy shift 0.015 against a 0.53 condition separation); eval-time transplant fidelity is unimpaired at maximum dose (KL 0.0135 vs. the dose-zero trio's 0.0170; the nearest-matched pair, dose-zero at epoch 290 against top dose at epoch 300, shows no cost either, KL 0.0124 vs. 0.0126); and ID accuracy is uncosted in the completed-schedule coordinate at every dose (0.8822–0.8858 vs. dose-zero reruns' 0.8786–0.8850); raw endpoint deltas include -2.1 and -2.3pp readings, but those are stop-timing artifacts rather than dose effects. One second-order co-movement is disclosed: at the top dose, corruption-stability shifts by -0.0043 ± 0.0020 (3/3 seeds negative, $\sim$6% of the condition separation, partly stop-timing), an order of magnitude smaller than the redundancy displacement.

Robustness does not follow. Table 1 is the pre-registered readout: matched-skill Δ effective robustness (dosed minus dose-zero, same seed) in the two registered coordinates, per-pair (each pair at its own matched level) and iso-level (all models at $L^* = 0.8248$). Every dose mean is inside $\pm$1pp in both coordinates: per-pair -0.29/+0.75/+0.23/-0.03pp, iso-level -0.57/+0.82/-0.37/-0.55pp. There is no monotone trend in either read, and per-seed directions split at $\lambda_{\text{adl}} \geq 1.0$. The two coordinates agree at the dose-mean level (within 0.61pp; seed 0's own level coincides with L^*, and seeds 1–2 differ between coordinates in opposite directions that cancel in the means). Seed 2's iso column is set throughout by one early-arriving baseline checkpoint at epoch 134, ID accuracy 0.8214 and effective robustness 0.480: the previous subsection's arrival heterogeneity expressed as a coordinate offset. Realized pair scatter ($\pm$1.7–3.8pp) is dominated by the same robustness-arrival heterogeneity as Section 4.3, so the verdict rests on dose means and two-coordinate agreement, not single pairs. Dose means are over three seeds with seed-level 95% intervals of $\pm$2.6–7.7pp against the 1pp bar: the design excludes a monotone effect at the 1pp scale, not seed-scale effects of a few pp, which is why the verdict is no detectable response rather than a bounded effect size.

Read against the verdict definitions frozen before these evaluations ran: **flat**. Forcing cross-row redundancy through half of the condition separation produced no detectable robustness response at matched skill, under either registered way of matching accuracy. This is consistent with the deficit residing in features, not in the visible routing. It does not prove the features account, since elimination plus intervention is not direct feature measurement, and its scope is this regime.

Secondary probes are in Appendix J: a frozen ImageNet-C panel that reproduces the maturity signature within-seed (+6.1pp with zero structural change) but on which the natural-shift ordering does *not* hold (the single fine-tuned and single scratch seeds both score above the completed distilled cluster; reported, not resolved), Q/K-projection stability, and the per-suite decomposition of the residual gap.

5 DISCUSSION

Three results triangulate one conclusion. The copy is verified near-perfect and training-invariant, so the deficit is not a degraded transfer. The gap closes within-seed under continued training while the transferred structure never moves, so what was missing matures slowly and is not the routing. And forcing the routing through half a condition-width of structural change, with accuracy, sharpness, and fidelity held, moves robustness by nothing detectable. The account left standing is Li et al.'s own untested conjecture, now with instrumentation behind it: attention transfer hands the student

Table 1: The causal test: matched-skill Δ effective robustness (dosed − dose-zero, same seed) per dose, in both registered coordinates. Read against the frozen verdict definitions: flat. Seed 0's two blocks repeat one measurement by construction; no bootstrap intervals exist for trajectory checkpoints (Section 3).

	Per-pair				Iso-level			
Dose	s0	s1	s2	mean	s0	s1	s2	mean
$\lambda_{\text{adl}} = 0.1$	+0.0154	+0.0017	-0.0259	-0.0029	+0.0154	+0.0054	-0.0380	-0.0057
$\lambda_{\text{adl}} = 0.3$	+0.0196	+0.0010	+0.0019	+0.0075	+0.0196	+0.0180	-0.0132	+0.0082
$\lambda_{\text{adl}} = 1.0$	+0.0236	-0.0176	+0.0010	+0.0023	+0.0236	+0.0025	-0.0373	-0.0037
$\lambda_{\text{adl}} = 3.0$	+0.0166	-0.0163	-0.0013	-0.0003	+0.0166	+0.0039	-0.0370	-0.0055

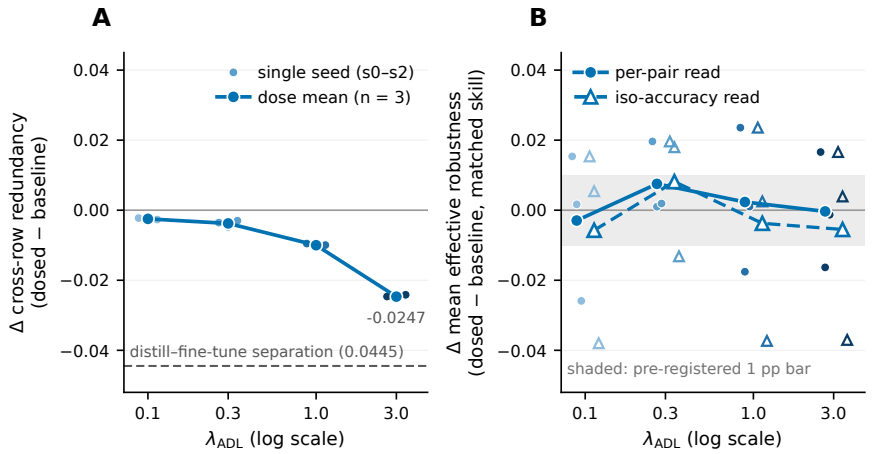

Figure 2: The causal test. **A:** paired same-seed change in cross-row redundancy per dose; the top dose displaces structure by half the distill–fine-tune separation (dashed line: magnitude ruler only; fine-tuning lies in the opposite direction, $+0.0445$). **B:** matched-skill change in mean effective robustness, per-pair (solid) and iso-accuracy (dashed), same vertical scale as A; shaded band = the pre-registered 1pp bar. The dial sweeps half a condition-width of structure, and robustness does not follow.

the teacher's routing but not the teacher's features, and the robustness the student eventually gains, it grows itself, slowly, under the transplanted gaze. We state the boundary as plainly as the finding: this is elimination plus intervention, not a direct exhibition of the features, and its scope is the regime we measured.

Our results join the "attention is not explanation" debate (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019) from a new axis: not whether attention explains predictions, but whether attention structure carries robustness. We ran the transplant experiment the correlational thread could not: hold the "good-looking" attention fixed by construction, then watch whether the property it supposedly certifies arrives. It does not. The descriptive use of attention maps survives untouched; the evidential use loses its cleanest support. In our regime, the student wearing the teacher's exact, diverse, corruption-stable gaze is less robust at rule-governed endpoints than the fine-tuned student whose attention drifted far from the teacher's. This is the reverse of the ordering attention quality would predict. The scratch student, whose attention drifted farthest of all, is nonetheless the least robust. Across conditions the two are simply not coupled. Attention overlays show where a model looks, not what it knows.

For practitioners, the deficit is, at our scale, substantially a budget artifact: distillation's robustness matures later than its accuracy, so a stopping rule tuned to accuracy flatness samples distilled models at their least robust moment. Three cheap corollaries: budget distillation schedules generously; never compare methods' robustness at a fixed budget without checking both maturity clocks; do not select checkpoints by attention aesthetics.

Our magnitudes are scale-bound, but the mechanism makes a testable claim at Li et al.'s scale: if ViT-L attention-transfer students at 200 epochs sit early on their own robustness-maturity curve, then extending distillation budgets there should compress the gap toward fine-tuning, with attention structure static throughout; if it instead persists at true completion, the maturity account is bounded to small models and the feature story needs a scale-dependent component. Either outcome is informative, and the experiment is nothing more than a budget extension.

One observation invites theory: supervised classification, left free, made attention *more* redundant in every condition that could drift, while reconstruction pretraining built the diverse, corruption-stable structure everyone admires. Whatever classification wants from attention, it is not what reconstruction builds; and whatever robustness needs, our intervention shows it is not obtained by pushing the redundancy score back down. Both halves are consistent with the features carrying the load, with attention structure as an expressive but non-binding byproduct: readable, transplantable, and causally light.

6 Limitations and design retrospective

Scale. Everything here is ViT-S on ImageNet-100 with an MAE-S teacher; magnitudes are regime-bound and claimed for no other setting. The mechanism-level claims travel on arguments, not evidence: the KL term binds routing as an optimization constraint rather than an emergent phenomenon, and portability remains argued, not shown. Our teacher and student are architecture-matched; attention transfer itself can fail under architectural mismatch (Qin et al., 2026), and our claims live in the matched regime, precisely where the robustness question is well-posed. The compensation for scale is resolution: the full-trajectory instrumentation every analysis rests on costs several hundred dollars at this scale, and roughly two orders of magnitude more at Li et al.'s. We bought density with size, and the exported prediction (Section 5) is the test we could not afford to run.

The stopping rule. Our early-stopping rule, standard patience on validation accuracy, turned out to be a peak-detector deployed on curves that have no peak: the overfit turn never arrives, and the rule's firing time becomes a lottery, stopping same-condition runs 145 epochs apart. We quantify it (Appendix E) and ground convergence claims in schedule completion and measured interruption costs rather than the rule's verdicts. On no-peak curves, complete the horizon and use stopping rules as monitors; pilot novel conditions at $n \geq 2$ seeds, since stop-variance is invisible at $n = 1$.

The protocol accumulated amendments mid-study; Appendix A lists the full set with their triggers, including one that is post-hoc for the two runs that prompted it and prospective for every case it subsequently governs. We regard such amendment as the correct response to an instrument failing in use, and document the timeline so reviewers can regard it otherwise if they choose.

Inference boundaries. The features account rests on elimination and intervention, and is consistent-with rather than proven: we do not exhibit the deficient features themselves. Three seeds per condition bound our variance estimates; one dataset, one teacher, one architecture family bound everything else. Bootstrap intervals quantify image-sampling noise, not run-to-run variation, which the seed bands carry. The effective-robustness definition is a boundary too: our ratio form and the fitted-baseline residual of Taori et al. (2020) disagree wherever compared models differ in ID accuracy. The verdicts rest on matched and near-equal-ID comparisons, but the maturity gradient's sign is definition-dependent (Appendix G).

AI use statement

We used LLM assistance to implement the training, evaluation, and analysis code, to provide feedback on experimental design, to assist in analyzing results and in literature search, and to draft and edit the manuscript and figures. All research decisions are the authors' own: the hypotheses, the registered predictions, the thresholds, and the interpretations. All LLM outputs were reviewed and verified by the authors; representative verification checks appear in Appendices B and D. We take full responsibility for the final content of this work.

Reproducibility statement

The full training configuration and compute accounting are in Appendix H; the distillation loss is verified against the reference implementation to float precision (Appendix B); the ImageNet-C panel is generated by a deterministic reference implementation and frozen (byte-identical across all evaluations); and continuation runs were validated by exact replay of their canonical prefixes across 131 shared evaluations (Appendix D). Appendix A lists the complete set of protocol commitments and amendments, with each amendment's trigger and its timing relative to the data it governed.

References

Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *Transactions on Machine Learning Research*, 2022. arXiv:2106.15831.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. arXiv:2011.03395.

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

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.11929.

Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17557–17568, 2023. arXiv:2303.11126.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2111.06377.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1903.12261.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021a. arXiv:2006.16241.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b. arXiv:1907.07174.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, 2019. arXiv:1902.10186.

Alexander C. Li, Yuandong Tian, Beidi Chen, Deepak Pathak, and Xinlei Chen. On the surprising effectiveness of attention transfer for vision transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2411.09702.

Kang Lin and Reinhard Heckel. Robustness of deep learning for accelerated MRI: Benefits of diverse training data. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2312.10271.

John P. Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2107.04649.

Armand Mihai Nicolicioiu, Andrei Liviu Nicolicioiu, Bogdan Alexe, and Damien Teney. Learning diverse features in vision transformers for improved generalization. *ICML Workshop on Spurious Correlations, Invariance and Stability*, 2023. arXiv:2308.16274.

Utkarsh Ojha, Yuheng Li, Anirudh Sundara Rajan, Yingyu Liang, and Yong Jae Lee. What knowledge gets distilled in knowledge distillation? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2205.16004.

Huaiyuan Qin, Muli Yang, Gabriel James Goenawan, Peng Hu, Chen Gong, Xi Peng, and Hongyuan Zhu. Attention transfer is not universally effective for vision transformers. *arXiv preprint arXiv:2605.07191*, 2026.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019. arXiv:1902.10811.

Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, 2020. arXiv:2002.11569.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.05945.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2007.00644.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2012.12877.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1905.13549.

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, Kai Wang, and Yang You. REPA works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.16792.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, 2019. arXiv:1908.04626.

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 (CVPR)*, 2024. arXiv:2404.01415.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1612.03928.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2303.06296.

Songming Zhang, Yuxiao Luo, Ziyu Lyu, and Xiaofeng Chen. ShiftKD: Benchmarking knowledge distillation under distribution shift. *Neural Networks*, 192:107838, 2025. arXiv:2312.16242.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Anima Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2204.12451.

A PROTOCOL COMMITMENTS AND AMENDMENTS

Fixed before any training run. The following were set before any compute was spent:

- the three training conditions and three seeds per condition;
- the single early-stopping rule of Section 3 and the per-condition cosine horizons;
- the effective-robustness gap threshold of 0.01, baked into the analysis code;
- the three hypotheses stated in Section 3;
- both comparison coordinates, endpoint and matched-accuracy, present in the analysis code before any result existed;
- the $\lambda_{\mathrm{kl}} = 36$ setting, from the derivation of Appendix B alone;
- the teacher quality gate, its mechanism fixed in advance and its floor pinned at launch from the measured probe.

Amended during the study. Each amendment below responded to an instrument behaving unexpectedly, and each governed only data that did not yet exist when it was adopted, with the one exception noted.

- *Fine-tune horizon, 100 to 200 epochs.* The initial 100-epoch schedule bound while validation was still climbing; the pre-declared pilot-calibration clause was exercised once, before the full matrix launched, and never again.
- *Cap-hit flatness reading* (net validation gain over the final 25 epochs against a 0.5pp bar). Adopted after the two fine-tune seed-0 cap-hits that prompted it, so it is post-hoc for those two runs and prospective for every later cap-hit (Appendix C).
- *Convergence-reporting policy* (stopping-rule verdicts are operational only; convergence claims rest on completed anneals, shown flatness, seed replication, and measured interruption cost). Adopted immediately after the study's first flatness-test failure, before any subsequent one.
- *Stop-invariant readout for the causal test* (matched-accuracy primary, with the iso-level variant). Registered after the baseline stop-timing variance was measured and before any dosed run existed; the iso-level specification was added before any dosed evaluation.
- *Full-schedule rerun program.* Registered before the matrix completed, with an escalation criterion (extend beyond seed 0 iff the seed-0 delta reached 1.0pp ID or 0.01 effective robustness) fixed before the first rerun existed; extended to all three seeds before any rerun data existed, superseding the criterion, which the seed-0 result then also satisfied.
- *Wave-2 ImageNet-C extension.* The additional role-labeled rows of Table 10 were registered before those evaluations existed, and extended to all three rerun endpoints when the rerun program grew.
- *Verdict-interpretation table.* The outcome-to-conclusion mapping for both verdicts was frozen before any verdict-bearing data existed; Sections 4.2 and 4.4 report their results against it.

One in-place correction. The registration record was edited in place exactly once: the note accepting the $\lambda 1.0$ s1 cap-hit described that run's trajectory tail incorrectly, and the description was corrected and labeled at the site (Appendix C carries the corrected numbers). Every other change over the study's life was an addition, and no rule, threshold, or reported result was ever edited. The registry and the project's working records are retained; this appendix is the complete list of protocol-relevant commitments and changes.

The second hypothesis. The prediction that the layer-wise overfocusing signature would predict the ID-to-OOD gap is reported nowhere as a fitted correlation, and the reason is recorded here. The

measurements left the registered form unanswerable rather than answered: attention structure proved near-invariant within condition (Section 4.1) while the gap proved highly variable with stopping time (Section 4.2), so an observational fit would relate a predictor taking three effective values, one per condition and collinear with it, to an outcome dominated by training maturity. The dose ladder of Section 4.4 tests the same linkage with the predictor manipulated within condition instead, and Section 5 reports the cross-condition observation in prose: the two are not coupled.

B THE $\lambda_{\text{kl}} = 36$ DERIVATION AND GRADIENT-IDENTITY CHECK

Li et al.'s released implementation computes the attention-transfer loss as a teacher-as-target cross-entropy over attention maps: gradient-identical to a KL divergence, since the teacher-entropy term is constant in the student, though offset in value. It is averaged over batch, heads, and query rows, then summed over the distilled layers, scaled by $\lambda = 3$. Our implementation averages over layers as well, a mean of means, which changes the effective per-layer weight. Matching Li et al.'s per-layer gradient scale at 12 distilled layers therefore requires $\lambda_{\text{kl}} = 3 \times 12 = 36$. This is arithmetic, not tuning: the two formulations are identical up to the constant. We verified the identity directly. Autograd gradients of the two loss formulations, $36\times$ mean-over-12-layers KL against $3\times$ the sum over 12 layers of Li et al.'s verbatim loss, were evaluated on identical randomly-drawn attention logits and agree to a maximum absolute difference of 5.6×10^{-9}, which is float32 noise. This is a loss-level gradient identity on synthetic inputs, which is the quantity the derivation asserts.

Two deliberate loss-side departures from Li et al.'s distillation recipe are disclosed here. Training-protocol differences such as budgets, stopping rule, and teacher provenance are Section 3's subject. First, we distill all 12 layers rather than their default first-18-of-24, for treatment purity: the analyses then measure the same quantity the loss optimizes, at every layer. Second, our batch size and learning-rate schedule follow the standard DeiT-style recipe at our scale rather than their ViT-L configuration. Our design rule was to be faithful to the method, not the numbers. The $\lambda_{\text{kl}} = 36$ value was pinned before the matrix ran, from this derivation alone; the pilot's transfer-bound check (the KL term collapsing $1.67 \to 0.02$ and flattening) confirmed it binds as intended.

C CONVERGENCE ADJUDICATION RECORDS

Every run ends in one of three recorded states: early-stop certificate (the rule fired), cap-hit with a passing terminal-flatness reading, or cap-hit with a failing one, which is never auto-accepted and always adjudicated on the record. The flatness reading is the net validation top-1 gain over the final 25 epochs against a 0.5pp bar; it is descriptive evidence, never a convergence certificate (per-eval noise is $\pm$0.44pp, so single-interval deltas are sub-noise by construction). Three cap-hits failed the bar and were accepted with disclosure. All three overages are small, 0.02–0.08pp over the 0.5pp bar, on fully-annealed tails whose per-eval gaps sit at or below the $\pm$0.44pp noise. None shows the consecutive above-noise climb of the one trend-shaped case observed, at +0.205pp/eval. The discriminator separates the classes by roughly $2\times$ in rate and 4–8$\times$ in overage.

- **finetune s2**: net +0.52pp vs. the 0.50 bar (fail by 0.02pp, one image); final five gaps +0.12/+0.12/+0.24/+0.06/$-$0.02, each at or below per-eval noise. Accepted under the completed-schedule policy with family-band evidence: the three fine-tuned seeds ended via three different formal outcomes (a rule-issued stop certificate, a certificate completed only at the horizon's final evaluation, and this failed flatness reading) yet landed in a 0.3pp ID band (0.8896–0.8926), effective robustness 0.4913–0.4964. The formalism's verdicts are demonstrably uncorrelated with the answers.
- λ**0.3 s0**: net +0.58pp (fail by 0.08); final gaps +0.16/+0.02/+0.38/$-$0.30/+0.32, the largest at epoch 289. Accepted per the same precedent.
- λ**1.0 s1**: net +0.54pp (fail by 0.04); final gaps +0.24/+0.14/+0.24/$-$0.12/+0.04, three positive: the least clean of the three accepts. Accepted on the quantitative discriminator (rate 0.108pp/eval, roughly half the one trend-shaped case ever observed; overage an eighth of that case's).

The discriminator separating noise-shaped from trend-shaped overage was validated by the one trend-shaped case: the original 100-epoch fine-tune horizon failed with +0.205pp/eval while still

climbing, and its extension recovered +2.0pp (0.8718 → 0.8916 best validation top-1), retro-verifying the classification. The false-fail mode of the flatness test is single-jitter aliasing, quantified above. Full validation trajectories for every run are retained with the run artifacts.

D CONTINUATION PROGRAM (FULL-SCHEDULE RERUNS)

Registration. The seed-0 rerun was registered before the matrix completed, with a pinned escalation criterion fixed before the run existed: extend to seeds 1–2 iff the seed-0 delta reached 1.0pp ID or 0.01 effective robustness. The criterion fired, at +6.0pp ID and +0.053 effective robustness. Before any continuation data existed the program had been amended to all three seeds regardless (Appendix A). That made seed 2 an in-design negative control with a predicted near-zero delta, since it had been cut only 10 epochs from its horizon, at 1.3% of peak learning rate.

Instrument validation. Same-seed replay is exact in this stack: across 131 shared evaluations spanning the three seeds, the reruns reproduce their canonical prefixes with a maximum absolute validation difference of 0.0. Each rerun is therefore the canonical run continued, not a new draw. The control seed moved +0.28pp ID / +0.71pp effective robustness, below the 1pp instrument-suspicion bar, and its nonzero remainder is the genuinely remaining 10 epochs of schedule.

Readings. Interruption cost is monotone in how hot the rule cut: 59% of peak learning rate at the cut recovered +6.02pp ID / +5.31pp effective robustness; 20% recovered +1.76 / +3.11pp; 1.3% of peak (the schedule's floor plus 0.3% of remaining decay) recovered +0.28 / +0.71pp. Effective robustness outgained ID accuracy in the two later seeds' windows; seed 0's full remainder moved both together (+5.31 vs. +6.02pp). Structure at the completed schedules' promoted checkpoints stayed at the teacher's value in all three seeds (redundancy 0.4673/0.4676/0.4673 against the teacher's 0.4665). Continuations are labeled artifacts with first-class evidential use and second-class authority: registered claims are computed from rule-endpoints alone, and no continuation number enters a primary table unlabeled.

E STOP-LOTTERY STATISTICS

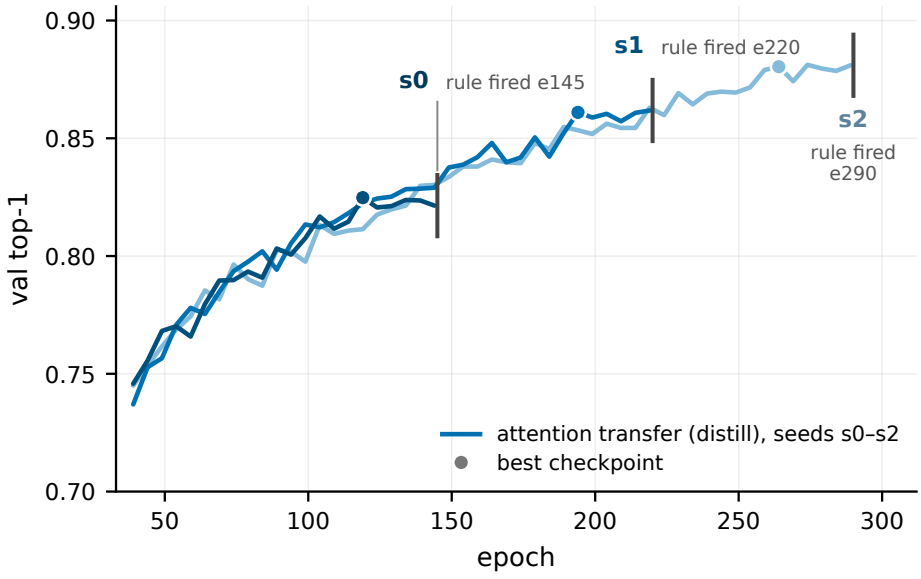

Figure 3: Validation trajectories of the three distilled seeds under the single pre-registered stopping rule, with rule-fire epochs and best checkpoints marked; line shade distinguishes seeds. Near-identical no-peak curves; stops 145 epochs apart.

The distill family exposes the stopping rule's variance. Under one rule and one configuration, three seeds stopped at epochs 145, 220, and 290 (Figure 3), giving an ID spread of 5.6pp and an effective-robustness spread of 4.1pp at the endpoints. The attention-structure metrics meanwhile replicated tightly across the same runs: redundancy ± 0.0008, top-k ± 0.0011, entropy ± 0.006, and stability ± 0.002, each one to three orders of magnitude below its between-condition separation. The curves themselves are near-identical: seed 1 at epoch 144 sat at 0.8290, within noise of seed 0's final 0.8248; the difference is where 25 flat epochs happened to align with the ± 0.44pp per-eval noise. The rule is a peak-detector and these curves have no peak, so its firing time is a lottery over noise alignments, and endpoint metrics inherit that lottery. Section 4.2's $r = 0.978$ is the same fact viewed across the family. Fine-tuning and scratch, whose plateaus fit within their horizons, show the ordinary version: seed bands of 0.3–0.5pp on both metrics. Fine-tuning spans ID 0.8896–0.8926 and effective robustness 0.4913–0.4964; scratch spans 0.8462–0.8498 and 0.4168–0.4199.

F Full per-suite tables and matched-pair selections

Tables 2, 3, 5, and 4 reproduce the endpoint, gap, matched-ladder, and dose tables referenced from the main text.

Table 6 reports per-suite effective robustness for all 24 endpoint models with bootstrap CIs on the 4-suite mean. Table 7 lists every matched-pair checkpoint selection used in Sections 4.3 and 4.4: run, target level, selected epoch, achieved validation accuracy, and residual. Residuals are at most 0.0034 for every dose and iso-level selection; the one exception is the seed-1 own-level ladder rung, whose fine-tuned side sits 0.0054 below target and is machine-flagged loose against the 0.005 tolerance. Within-pair ID differences are at most 0.22pp in the dose analyses, 0.36pp in the iso-level analyses, and 0.54pp at that one ladder rung.

Table 2: Rule-governed endpoints per condition and seed, with labeled full-schedule reruns. Effective robustness is the 4-suite mean; brackets are paired image-level bootstrap 95% CIs (B=10,000). The evaluated checkpoint is each run's min-delta-promoted best, which may precede the stop epoch shown.

Condition	Seed	Stop epoch	ID top-1	Eff. robustness [95% CI]
Scratch	s0	300 (cap)	0.8498	0.4199 [0.4070, 0.4332]
	s1	300 (cap)	0.8496	0.4168 [0.4038, 0.4301]
	s2	290 (early stop)	0.8462	0.4188 [0.4053, 0.4324]
Weight transfer (ft)	s0	200 (cap)	0.8916	0.4943 [0.4814, 0.5073]
	s1	200 (early stop)	0.8896	0.4913 [0.4785, 0.5044]
	s2	200 (cap)	0.8926	0.4964 [0.4836, 0.5092]
Attention transfer (distill)	s0	145 (early stop)	0.8248	0.4369 [0.4233, 0.4510]
	s1	220 (early stop)	0.8610	0.4590 [0.4461, 0.4720]
	s2	290 (early stop)	0.8804	0.4783 [0.4658, 0.4910]
distill, full-schedule rerun	s0	300 (no rule)	0.8850	0.4900 [0.4773, 0.5023]
	s1	300 (no rule)	0.8786	0.4901 [0.4774, 0.5030]
	s2	300 (no rule)	0.8832	0.4854 [0.4727, 0.4980]

Table 3: The robustness gap (ft − distill, mean effective robustness) per seed, in the endpoint coordinate (rule-governed stops) and at completed schedules (full-schedule reruns). Pre-registered threshold 0.01. Paired bootstrap 95% CIs.

Seed	Endpoint gap	Completed-schedule gap
s0	+0.0575 [0.0445, 0.0703]	+0.0044 [-0.0059, 0.0148]
s1	+0.0323 [0.0209, 0.0442]	+0.0013 [-0.0096, 0.0122]
s2	+0.0181 [0.0073, 0.0289]	+0.0110 [0.0003, 0.0217]

Table 4: The causal dial: paired same-seed change in cross-row redundancy at each ADL dose (sweep endpoint minus its dose-zero baseline). Negative = less redundant.

Dose	s0	s1	s2	Mean ± SD
$\lambda_{\text{adl}} = 0.1$	-0.0022	-0.0025	-0.0027	-0.0025 ± 0.0002
$\lambda_{\text{adl}} = 0.3$	-0.0035	-0.0048	-0.0030	-0.0038 ± 0.0009
$\lambda_{\text{adl}} = 1.0$	-0.0095	-0.0105	-0.0100	-0.0100 ± 0.0005
$\lambda_{\text{adl}} = 3.0$	-0.0246	-0.0252	-0.0241	-0.0247 ± 0.0006

Table 5: The matched-accuracy ladder (wave-2): gap (ft − distill, mean effective robustness) at each seed's own matched level and at the top matched level, trajectory checkpoints both sides; top-rung distill checkpoints come from the labeled full-schedule reruns. Selection is by nearest ID to the target level; for s1's top rung this picks the rerun's epoch-299 checkpoint rather than its promoted best, and the gap stays inside the band under either choice (Section 4.3). Bootstrap intervals do not exist for trajectory checkpoints (Section 3).

	Own-level rung		Top rung	
Seed	level	gap	level	gap
s0	0.8248	+0.0237	0.8850	+0.0037
s1	0.8618	-0.0145	0.8794	-0.0030
s2	0.8812	+0.0049	0.8832	+0.0049

Table 6: Per-suite effective robustness for every evaluated endpoint model. Mean column carries the paired bootstrap 95% CI.

Model	IN-A	IN-R	IN-Sk	IN-V2	Mean [95% CI]
scratch_vit_s_s0	0.1088	0.3871	0.3199	0.8637	0.4199 [0.4070, 0.4332]
scratch_vit_s_s1	0.1196	0.3919	0.3107	0.8451	0.4168 [0.4038, 0.4301]
scratch_vit_s_s2	0.1303	0.3922	0.3122	0.8402	0.4188 [0.4053, 0.4324]
finetune_vit_s_s0	0.1453	0.4953	0.4405	0.8961	0.4943 [0.4814, 0.5073]
finetune_vit_s_s1	0.1564	0.4805	0.4302	0.8982	0.4913 [0.4785, 0.5044]
finetune_vit_s_s2	0.1433	0.4965	0.4440	0.9019	0.4964 [0.4836, 0.5092]
distill_vit_s_s0	0.0829	0.4195	0.3964	0.8487	0.4369 [0.4233, 0.4510]
distill_vit_s_s1	0.0913	0.4635	0.4101	0.8711	0.4590 [0.4461, 0.4720]
distill_vit_s_s2	0.1034	0.4872	0.4492	0.8735	0.4783 [0.4658, 0.4910]
distill_vit_s_s0_fullsched	0.1204	0.4998	0.4447	0.8949	0.4900 [0.4773, 0.5023]
distill_vit_s_s1_fullsched	0.1184	0.5023	0.4473	0.8923	0.4901 [0.4774, 0.5030]
distill_vit_s_s2_fullsched	0.1137	0.4972	0.4409	0.8899	0.4854 [0.4727, 0.4980]
distill_h3_vit_s_lam0.1_s0	0.0876	0.4349	0.3886	0.8464	0.4394 [0.4261, 0.4532]
distill_h3_vit_s_lam0.1_s1	0.1227	0.4983	0.4461	0.8932	0.4901 [0.4772, 0.5032]
distill_h3_vit_s_lam0.1_s2	0.0887	0.4471	0.3897	0.8725	0.4495 [0.4367, 0.4628]
distill_h3_vit_s_lam0.3_s0	0.1127	0.4942	0.4474	0.8885	0.4857 [0.4731, 0.4984]
distill_h3_vit_s_lam0.3_s1	0.1182	0.4916	0.4431	0.8829	0.4840 [0.4713, 0.4967]
distill_h3_vit_s_lam0.3_s2	0.1200	0.5022	0.4546	0.8954	0.4931 [0.4805, 0.5061]
distill_h3_vit_s_lam1.0_s0	0.1260	0.4836	0.4360	0.9024	0.4870 [0.4743, 0.5000]
distill_h3_vit_s_lam1.0_s1	0.1229	0.4871	0.4526	0.8857	0.4871 [0.4743, 0.4999]
distill_h3_vit_s_lam1.0_s2	0.0969	0.4529	0.4064	0.8897	0.4615 [0.4486, 0.4746]
distill_h3_vit_s_lam3.0_s0	0.1048	0.4923	0.4465	0.8698	0.4783 [0.4656, 0.4911]
distill_h3_vit_s_lam3.0_s1	0.1056	0.4853	0.4486	0.8806	0.4800 [0.4674, 0.4929]
distill_h3_vit_s_lam3.0_s2	0.1211	0.5016	0.4527	0.8863	0.4904 [0.4777, 0.5034]

Table 7: Every matched-pair checkpoint selection: analysis, model, achieved validation top-1, selected epoch, and residual to the target level. All residuals ≤ 0.0034 except the s1 own-level ladder rung's fine-tuned side (-0.0054, machine-flagged loose against the 0.005 tolerance).

Analysis	Model	Val top-1	Epoch	Residual
dose per-pair λ0.1	sweep s0	0.8234	129	-0.0014
dose per-pair λ0.1	baseline s0	0.8248	119	+0.0000
dose per-pair λ0.1	sweep s1	0.8622	219	+0.0004
dose per-pair λ0.1	baseline s1	0.8618	219	+0.0000
dose per-pair λ0.1	sweep s2	0.8596	189	+0.0000
dose per-pair λ0.1	baseline s2	0.8598	224	+0.0002
dose per-pair λ0.3	sweep s0	0.8242	139	-0.0006
dose per-pair λ0.3	baseline s0	0.8248	119	+0.0000
dose per-pair λ0.3	sweep s1	0.8612	219	-0.0006
dose per-pair λ0.3	baseline s1	0.8618	219	+0.0000
dose per-pair λ0.3	sweep s2	0.8818	274	+0.0006
dose per-pair λ0.3	baseline s2	0.8812	274	+0.0000
dose per-pair λ1.0	sweep s0	0.8228	129	-0.0020
dose per-pair λ1.0	baseline s0	0.8248	119	+0.0000
dose per-pair λ1.0	sweep s1	0.8622	209	+0.0004
dose per-pair λ1.0	baseline s1	0.8618	219	+0.0000
dose per-pair λ1.0	sweep s2	0.8582	209	+0.0000
dose per-pair λ1.0	baseline s2	0.8598	224	+0.0016
dose per-pair λ3.0	sweep s0	0.8246	144	-0.0002
dose per-pair λ3.0	baseline s0	0.8248	119	+0.0000
dose per-pair λ3.0	sweep s1	0.8596	224	-0.0022
dose per-pair λ3.0	baseline s1	0.8618	219	+0.0000
dose per-pair λ3.0	sweep s2	0.8808	274	-0.0004
dose per-pair λ3.0	baseline s2	0.8812	274	+0.0000
iso-level ($L^* = 0.8248$)	sweep λ0.1 s0	0.8234	129	-0.0014
iso-level ($L^* = 0.8248$)	sweep λ0.1 s1	0.8230	124	-0.0018
iso-level ($L^* = 0.8248$)	sweep λ0.1 s2	0.8250	139	+0.0002
iso-level ($L^* = 0.8248$)	sweep λ0.3 s0	0.8242	139	-0.0006
iso-level ($L^* = 0.8248$)	sweep λ0.3 s1	0.8272	149	+0.0024
iso-level ($L^* = 0.8248$)	sweep λ0.3 s2	0.8242	134	-0.0006
iso-level ($L^* = 0.8248$)	sweep λ1.0 s0	0.8228	129	-0.0020
iso-level ($L^* = 0.8248$)	sweep λ1.0 s1	0.8248	134	+0.0000
iso-level ($L^* = 0.8248$)	sweep λ1.0 s2	0.8248	139	+0.0000
iso-level ($L^* = 0.8248$)	sweep λ3.0 s0	0.8246	144	-0.0002
iso-level ($L^* = 0.8248$)	sweep λ3.0 s1	0.8254	129	+0.0006
iso-level ($L^* = 0.8248$)	sweep λ3.0 s2	0.8234	124	-0.0014
iso-level ($L^* = 0.8248$)	baseline s0	0.8248	119	+0.0000
iso-level ($L^* = 0.8248$)	baseline s1	0.8244	124	-0.0004
iso-level ($L^* = 0.8248$)	baseline s2	0.8214	134	-0.0034
ladder top s0	distill side	0.8850	294	+0.0000
ladder top s0	ft side	0.8844	164	-0.0006
ladder top s1	distill side	0.8794	299	+0.0000
ladder top s1	ft side	0.8806	149	+0.0012
ladder top s2	distill side	0.8832	299	+0.0000
ladder top s2	ft side	0.8802	149	-0.0030
ladder own-level s0	distill side	0.8248	119	+0.0000
ladder own-level s0	ft side	0.8240	69	-0.0008
ladder own-level s1	distill side	0.8618	219	+0.0000
ladder own-level s1	ft side	0.8564	104	-0.0054
ladder own-level s2	distill side	0.8812	274	+0.0000
ladder own-level s2	ft side	0.8802	149	-0.0010

G RESIDUAL-FORM SENSITIVITY CHECK

We report effective robustness as the ratio OOD/ID on covered classes (Section 3); Taori et al. (2020) and Miller et al. (2021) define it as the residual of OOD accuracy above a baseline fit to ID accuracy in a scaled space (log-linear in Taori et al., probit in Miller et al.; our check follows Miller). The two definitions order models identically wherever compared models sit at matched or near-equal ID, which is where both of this paper's verdicts live. They can disagree where ID differs, which covers two reported quantities: the endpoint-coordinate gap triple and the maturity correlation. Table 8 recomputes both under the residual form for three baseline-population choices. This check is post-hoc: it was computed at final review and is not part of the pre-registered analysis set.

Table 8: Residual-form sensitivity check (post-hoc, computed at final review; not pre-registered). Per suite, probit(OOD) is fit linearly on probit(covered-ID) over the stated baseline population; a model's score is its mean residual over the four suites (probit units). Ratio-form references: $r = +0.978$; endpoint gaps +0.0575 / +0.0323 / +0.0181.

Baseline population	r(stop, residual)	ft − distill endpoint residual gap		
		s0	s1	s2
all 24 endpoint models	-0.562	-0.0181	-0.0122	+0.0525
fine-tune + scratch	-0.854	-0.0841	-0.0464	+0.0421
ft + scratch + dose-zero distill	-0.689	-0.0335	-0.0255	+0.0515

Under every non-degenerate baseline population the residual-form maturity correlation is negative, -0.56 to -0.85, where the ratio form reads +0.978. Later distilled checkpoints gain OOD accuracy, but more slowly than the cross-model probit trend predicts from their ID gains. The endpoint residual gaps also change sign for seeds 0 and 1. Neither movement touches the verdicts, which rest on matched- and near-equal-ID comparisons, but it means the maturity gradient and the endpoint triple are ratio-form statements, and readers in the residual tradition should read them as such (the main text labels them so). An in-study baseline of 24 models spanning 0.825–0.893 ID under-determines a Taori-style fit, so this is a sensitivity check, not a replacement measurement; the direction is nonetheless stable across every non-degenerate fit we tried.

H TRAINING CONFIGURATION AND COMPUTE

All runs share one training stack, resolved from version-controlled configuration files; the fully resolved configuration is stored within every run's artifacts. The classifier recipe is the standard DeiT-style setup at this scale: AdamW with weight decay 0.05, batch size 256, cosine learning-rate decay to a floor of 10^{-5}, gradient clipping at 1.0, label smoothing 0.1, stochastic depth 0.1, mixed-precision training, and the DeiT augmentation stack (RandAugment m9, mixup 0.8, cutmix 1.0, color jitter 0.3, random erasing 0.25). Validation and checkpointing run every 5 epochs (Section 3). Checkpoints are stored in full precision.

Table 9: Per-condition training settings; all other hyperparameters are shared as above.

	Peak LR	Warmup (epochs)	Horizon (epochs)
Scratch, distill, and dose arms	10^{-3}	20	300
Fine-tune	5×10^{-4}	5	200
MAE teacher (pretraining)	1.5×10^{-4}	40	400 (probe-stopped)

The MAE teacher additionally uses mask ratio 0.75 with a 256-dimensional, 4-layer decoder and normalized-pixel loss, decaying to a 10^{-6} floor. All runs used one NVIDIA A100-SXM4-40GB (PyTorch 2.12.1, CUDA 12.6, timm 1.0.27, Python 3.10); training compute, summed from per-run logs, was 282.5 GPU-hours over the 24 classifier runs (9 core, 12 dosed, 3 reruns) plus 8.7 for the teacher, about 291 in all.

I ADDITIONAL SETUP DETAILS

Teacher selection and quality gate. The MAE teacher was pretrained for up to 400 epochs with linear-probe-based early stopping (probes every 25 epochs from epoch 24, patience 3). The rule promoted the checkpoint with the best in-training probe, epoch 299; three later probes were higher by at most 0.0034, which the rule's 0.005 min-delta counts as no improvement. Because every downstream run depends on this one checkpoint, a quality gate guarded the spending against teacher regression or mis-loading: its mechanism was pre-registered, and its floor was pinned at launch as the teacher's measured probe accuracy minus a safety margin, 0.48 against a measured 0.5556. The two probe accuracies come from two instruments: the 0.5556 is a full 50-epoch linear head trained on the promoted checkpoint, while the in-training probes read by the stopping rule use a shorter head and sit lower, at 0.5434 for the promoted checkpoint.

Horizons in context. The horizons in context: the random-init family's 300-epoch horizon is, in hindsight, 2.07× the first observed distillation stop, though later seeds stopped as late as epoch 290; fine-tuning's doubled schedule landed at roughly twice its observed plateau epoch. The bootstrap is paired (one index draw per replicate, applied to every model) and propagates ID resamples into the covered-class denominators; the intervals quantify image-sampling variability on the fixed benchmarks, conditional on their class composition, with the covered-class sets held fixed as properties of the frozen benchmarks. The per-image outcome records backing the bootstrap cover the endpoint models only, so matched-coordinate quantities computed on trajectory checkpoints carry no bootstrap intervals; their uncertainty is carried by the single-pair noise scale (±0.5–1pp) and the disclosed seed spreads. The frozen ImageNet-C is self-generated from the ImageNet-100 validation set with the reference implementation (PNG, byte-identical across all conditions and evaluations). In the ADL objective, Guo et al.'s released implementation averages heads before the loss; we keep the per-head form of their Eq. 1–2 so the penalized quantity is identical to the measured one.

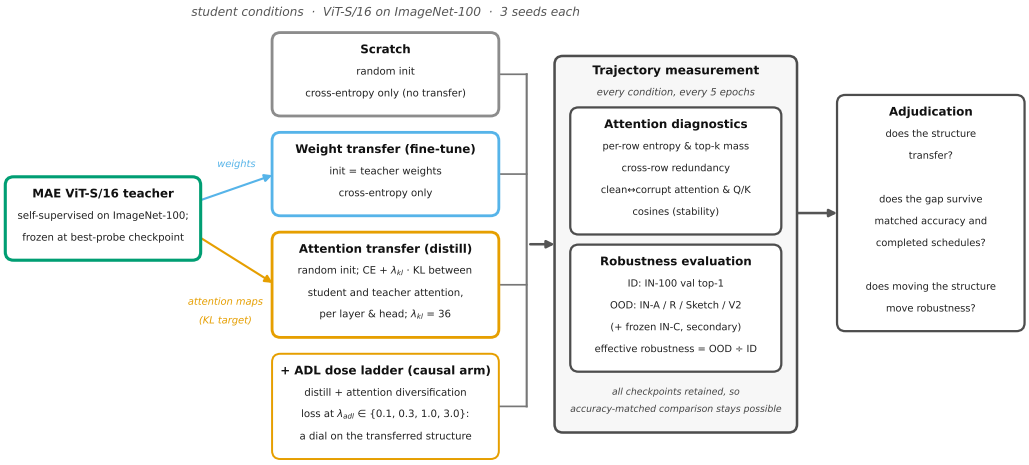

Figure 4: Study design: one probe-selected MAE teacher; three training conditions (scratch, weight transfer, attention transfer) × three seeds under one pre-registered stopping rule; an ADL dose ladder on the distillation objective; full-trajectory checkpoint retention feeding the matched-accuracy analyses; and labeled full-schedule reruns as the calibration layer.

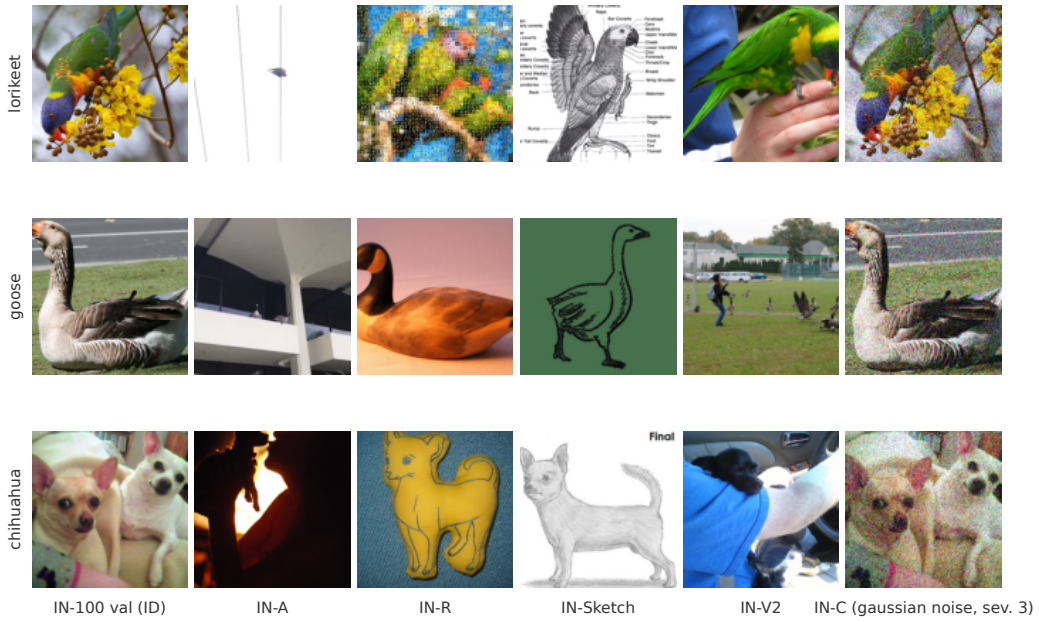

Figure 5: Three classes across the evaluation suites: the ID validation image and the same class as it appears in each natural-shift suite, plus an IN-C panel rendered from the ID image with the study's reference corruption implementation (gaussian noise, severity 3). Selection is by fixed rule, with no per-image curation: the classes are the first three, in sorted WordNet-ID order, of the IN-100 classes covered by both IN-A and IN-R, and each panel is the first file in sorted filename order in its class directory, shown under the evaluation crop.

J SECONDARY PROBES

Synthetic corruptions. ImageNet-C is our secondary, role-labeled probe (Table 10; frozen corruption set, identical bytes for every evaluation). It reproduces the maturity signature within-seed: the seed-0 distilled endpoint scores 0.6453 at its epoch-145 cut and 0.7067 after completing the schedule, a +6.1pp movement with zero structural change. The three completed distilled seeds cluster at 0.7067–0.7122, with the epoch-290 canonical s2 at 0.7052 as a second witness. The synthetic-shift picture differs from the natural-shift one. The single fine-tuned IN-C row (seed 0) scores 0.7548, above the completed distilled cluster by $\sim$4.5pp, and the single scratch row scores 0.7220, also above that cluster. On this probe the natural-shift ordering does not hold. We report the asymmetry rather than resolve it: it rests on one fine-tuned and one scratch seed, IN-C's corruption families interact with the training augmentations differently across conditions, and no primary claim cites IN-C alone.

Q/K projections under corruption. The stability ordering survives one level deeper than attention maps. Clean-corrupt cosines of the query and key projections are higher for distilled students, at Q 0.795 and K 0.760, than for fine-tuned students, at Q 0.733 and K 0.729, or scratch students, at Q 0.734 and K 0.736 (all seed means). We report this as directional evidence only: it locates the stability difference at or before the Q/K linear maps, and we do not pursue per-layer attribution.

Where the residual gap lives. The completed-schedule gap decomposes per suite. Fine-tuning leads on ImageNet-A at every seed, by +2.5/+3.8/+3.0pp; this is the suite with the widest paired-delta CIs. Completed distillation exceeds fine-tuning on ImageNet-R at all three seeds and on Sketch in two of three. ImageNet-V2 shows small fine-tune leads at every seed, +0.1 to +1.2pp. The residual completed-schedule gap is therefore not a uniform deficit: fine-tuning's lead concentrates in IN-A and V2, while distillation leads on R throughout and Sketch is split. Per-suite values for every model, with CIs, are in Appendix F.

Table 10: Secondary synthetic-shift probe: ImageNet-C mean effective robustness on the frozen corruption set, role-labeled. Registered canonical rows plus the registered wave-2 maturity extension.

Model	Role	IN-C eff. rob.
distill_s0 (e145 endpoint)	canonical	0.6453
finetune_s0	canonical	0.7548
scratch_s0	canonical	0.7220
distill+ADL λ1.0_s0	canonical	0.7101
distill_s0 continuation (e300)	maturity contrast	0.7067
distill_s1 continuation (e300)	maturity contrast	0.7098
distill_s2 continuation (e300)	maturity contrast	0.7122
distill_s2 (e290 endpoint)	second witness	0.7052
distill+ADL λ3.0_s0	dose probe	0.6948

K ADDITIONAL TABLES AND FIGURES

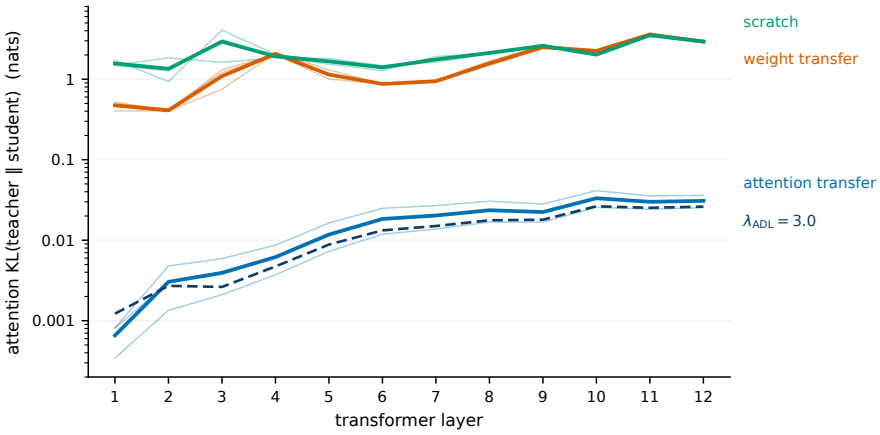

Figure 6: Per-layer attention KL(teacher||student) on held-out validation images (log scale; thin lines seeds, bold lines seed means). Distilled students sit roughly two orders of magnitude below both contrasts at every layer; maximum-dose ADL students (dashed) track the dose-zero band (mean 0.0135 vs. 0.0170 nats).

Table 11: Eval-time transplant fidelity on held-out validation images: mean per-layer/head KL(teacher||student) and cosine similarity of attention maps. Mean (min–max) over seeds.

Student	KL(t‖s) (nats)	Cosine	n
Attention transfer	0.0170 (0.0124–0.0217)	0.989 (0.986–0.992)	3
+ ADL λ=3.0	0.0135 (0.0126–0.0141)	0.991 (0.990–0.991)	3
Weight transfer	1.6545 (1.6243–1.6800)	0.499 (0.494–0.507)	3
Scratch	2.1482 (2.0881–2.2270)	0.342 (0.339–0.345)	3

Table 12: Attention structure at rule-governed endpoints (seed means; per-seed dispersion in Appendix E). Entropy and top-k mass measure per-row concentration; cross-row redundancy is Guo et al.'s diversity quantity (higher = rows more alike); stability is the clean–corrupt attention cosine.

Model	Entropy	Top-k	Redundancy	Stability
Teacher (MAE)	3.596	0.418	0.4665	–
Attention transfer	3.610	0.415	0.4683	0.810
Weight transfer	4.136	0.305	0.5128	0.742
Scratch	4.338	0.257	0.5837	0.750

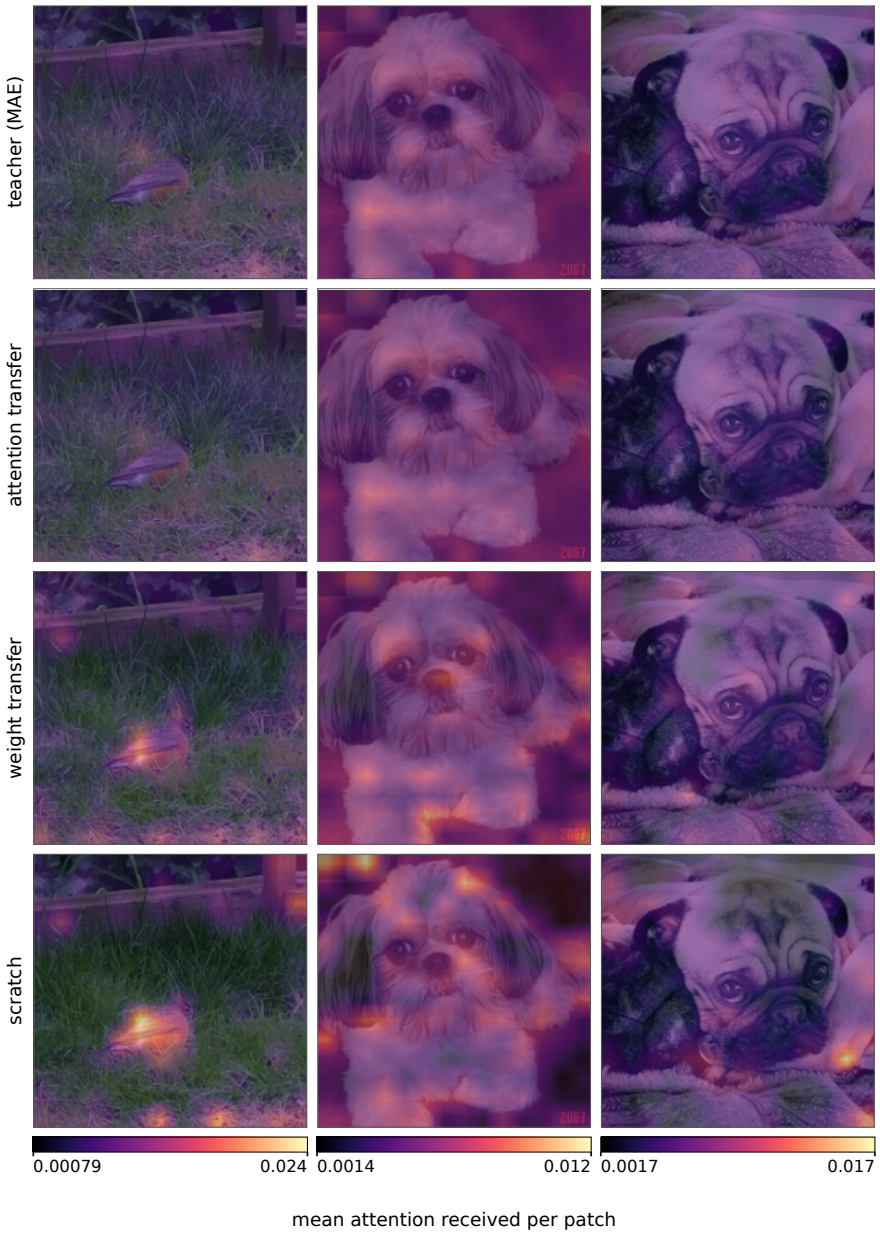

Figure 7: Mean attention received per patch (layer 6 of 12, averaged over heads and patch queries; CLS excluded as query), for three validation images: the first three of the study's fixed eight-image attention snapshot, with no per-image selection. The color scale is shared within each column, so panels are comparable across models; per-column ranges are printed below. The layer was selected by a fixed rule (the layer whose teacher–distill similarity is the median across layers), not for visual effect. The comparison is one of *similarity* only: the attention-transfer row reproduces the teacher's maps (cosine 0.998–0.999 on these images); weight transfer and scratch differ from both. Nothing in this figure bears on which attention is *better*; Section 4.4 addresses that question causally.

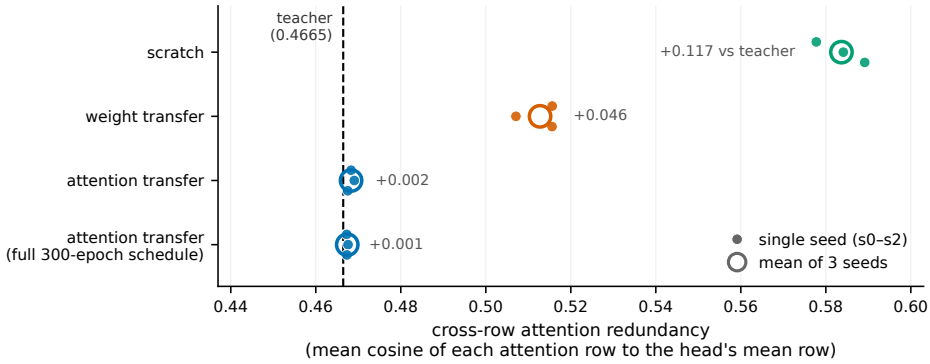

Figure 8: Cross-row redundancy at rule-governed endpoints (mean cosine of each thresholded attention row to the head's mean row, Guo Eq. 1–2; dots: seeds; open markers: seed means; dashed line: teacher). Distilled students sit at the teacher's value, including after full 300-epoch schedules; supervised training drifts redundancy upward when unconstrained.

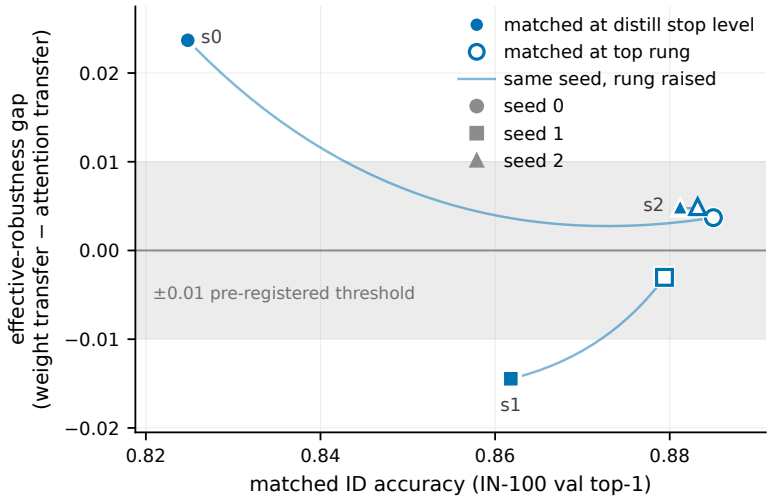

Figure 9: The matched-accuracy gap per seed at its own matched level (filled) and at the top matched level (open); connectors join each seed's two measured rungs and are annotations, not measured paths. Shaded band = pre-registered ±0.01 threshold. Gaps of both signs converge into the band as the matched level rises.