When More Adaptation Hurts: Degradation-Selective Sensitivity in CLIP Fine-tuning

Anonymous authors
Paper under double-blind review

Abstract

Fine-tuning on degraded data should improve robustness, but we find the opposite can occur. We uncover degradation-selective sensitivity in CLIP fine-tuning: robustness is not uniform, but depends jointly on adaptation depth L and learning rate η, an interaction concealed by aggregate accuracy. On a fine-grained bird benchmark with naturally occurring degradation, full-network adaptation at $\eta = 10^{-4}$ retains 98.5% accuracy on sharp images but drops to 78.8% on slight blur, while partial adaptation stays stable; across three domains and three seeds, high-dose full adaptation consistently enters a high-risk regime. Mechanistically, failure reflects how adaptation directions interact with decision geometry, not how far representations move: drift magnitude is matched across strata, and norm-matched counterfactuals show learned drift is protective yet less so on degraded inputs. A dose sweep shows that adaptation continues after error recovery saturates, while a paired within-domain control implicates directional motion degradation rather than blur magnitude; a minimal two-dimensional model illustrates how matched-margin samples can persist under late directional updates. We formalize this through an adaptation tolerance τ, which predicts co-adapted deployment-head errors (AUC 0.96 vs. 0.76 for a single-normal baseline). Degradation-stratified validation exposes these high-risk configurations.

1 Introduction

Fine-tuning vision-language models such as CLIP Radford et al. (2021) is a widely used way to adapt general-purpose representations to specialized tasks. The recipe—particularly learning rate η and adaptation depth L (how much of the backbone is unfrozen)—controls an *adaptation dose*, the extent and intensity of intervention on the pretrained representation. Such adaptation is typically tuned under an implicit trust that it is monotonically beneficial for the data being optimized on. We show this trust is not always warranted—at high LR, full-network adaptation can regress *below zero-shot* on the very degradation strata present in the training data. When training data contain heterogeneous visual degradation such as motion blur, defocus, or turbulence, aggressive adaptation can *selectively destroy* the competence it was meant to acquire—invisibly, because the loss is confined to one degradation stratum that neither aggregate top-1 nor sharp accuracy reveals; the failure is consequential yet easy to miss.

Discovery. Using CLIP ViT-B/32 as a controlled setting, full-network adaptation at $\eta{=}10^{-4}$ on a fine-grained bird task with naturally occurring degradation preserves sharp accuracy ($98.5 \pm 0.2\%$) but collapses slight-blur accuracy ($78.8 \pm 0.3\%$), a 19.7 ± 0.3pp gap, while partial adaptation (last 6 blocks) stays stable. Aggregate top-1 masks this as mild underperformance—we call it *degradation-selective sensitivity*. The sweep reveals an *adaptation-dose risk landscape*—under-fitting, beneficial adaptation, and high-risk failure—across depth and learning rate.

Replication and mechanism. Unlike LP-FT/WiSE-FT Kumar et al. (2022); Wortsman et al. (2022), which characterize an aggregate penalty under unseen test-time shift, we decompose the adaptation dose into stratum-specific effects. The Full–best-partial penalty grows with η and replicates across three domains and three seeds; the strongest selective phenotype is Bird-specific. The difficult step is attribution: the obvious account—"Full simply moves representations more"—is wrong, and ruling it out required frozen-boundary counterfactuals, norm-matched controls, and a formal tolerance measure. At the failing config drift magnitude is stratum-matched (≈ 1.0) and distributed (PC1

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.

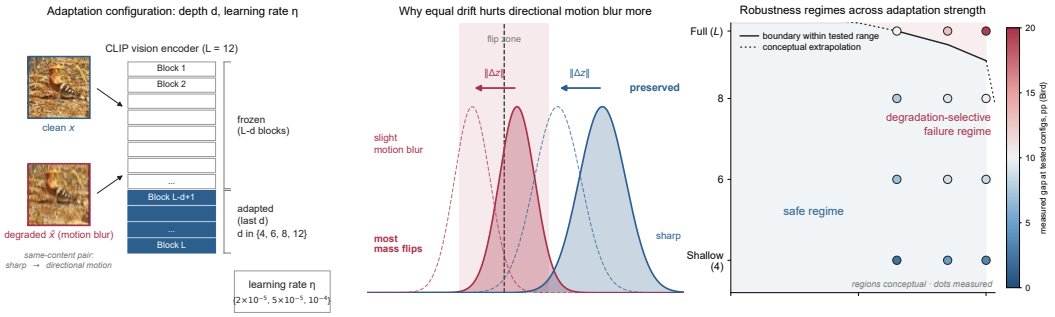


Figure 1: **Conceptual overview.** (a) Sharp vs. same-content directional-motion inputs under a depth×LR sweep. (b) Equal-magnitude drift (identical arrows) flips the slight stratum more because it starts closer to the boundary. (c) Risk landscape over depth and η: Full incurs the largest penalty at high η (dots measured; regions conceptual). Data in Fig. 2; details in Fig. 3.

$\approx 10\%$), so the explanation must be directional. Fixed-boundary counterfactuals show the learned direction is protective but 25pp more so for sharp than slight-blur inputs; an *adaptation tolerance* τ (margin per unit adverse drift) predicts independently co-adapted head errors (AUC 0.96). A dose sweep shows dose also determines *which* directions adapt (Sec. 6). A paired within-domain control reproduces the gap under directional motion blur (25.1pp) but not isotropic blur, locating selectivity in degradation geometry.

These findings resist straightforward remedies. Neither a learned depth router nor a semantic-preserving objective recovers robustness, because the defect is distributed, not localized to one fragile direction; degradation-stratified validation instead acts as a simple empirical screen (Sec. 7).

Why this matters. We ask why one adaptation dose can spare clean inputs while disproportionately harming a particular degradation stratum; the answer, a joint interaction of adaptation dose and decision-boundary tolerance, reframes robustness as stratified and geometric with a diagnostic (τ) and a stratified selection protocol. In one line: depth and LR determine what the representation learns, and per-stratum decision geometry determines what that change destroys.

Contributions.

1. *Two-tiered phenomenon.* We separate adaptation *risk* from degradation-*selective* failure: the depth–LR penalty (Full–partial -9.1 to -16.1pp at $\eta{=}10^{-4}$) replicates across fine-grained, coarse, and ImageNet-subset tasks, while selective collapse (98.5 $\rightarrow$ 78.8% on Bird) is conditional on task and degradation geometry rather than universal.
2. *Adaptation-dose view of fine-tuning risk.* Depth and optimization intensity jointly shape a dose-dependent gain–damage regime spanning under-fitting, beneficial adaptation, and degradation-sensitive failure: on ImageNet-100, L6 stays above zero-shot on every stratum ($+7.7$–$+25.2$pp) while Full regresses 8.4pp below zero-shot on slight blur at 10^{-4}.
3. *Directional mechanism.* Selective failure is not explained by drift magnitude: under norm-matched drift (22.3% vs. 5.8%, 3.8×), slight-blur samples undergo more prediction flips because the learned update interacts with per-stratum decision geometry—an effect paired same-content operators localize to directional-motion geometry, not blur magnitude.
4. *Predictive formalization.* Adaptation tolerance τ (margin per unit adverse drift) predicts independent co-adapted deployment-head errors (AUC 0.96 vs. 0.76).

2 Related Work

CLIP adaptation strategies. Methods for adapting CLIP Radford et al. (2021); Cherti et al. (2023) allocate trainable capacity through prompt learning Zhou et al. (2022b;a); Khattak et al. (2023); Jia et al. (2022), parameter-efficient modules Hu et al. (2022); Houlsby et al. (2019); Mishra et al. (2026); Li et al. (2024); Yamaguchi et al. (2025); Liu et al. (2025), or weight-space interpolation Wortsman et al. (2022). Those families sweep their capacity—rank, prompt length, module count—but that is a single scalar tuned under a fixed depth, not an interaction between depth and the

optimization regime, and none scores the result conditionally on degradation stratum. We instead hold the method fixed and treat the two levers every strategy implicitly commits to, the number of unfrozen blocks and the learning rate, as the variables of study, showing that their interaction is what determines degradation-stratified robustness.

VLM robustness under distribution shift. A large body of work characterizes the frozen CLIP model—its behavior under corruptions Taori et al. (2020); Hendrycks & Dietterich (2019); Wang et al. (2024), adversarial attacks Schlarmann et al. (2024), and natural shift, together with persistent shortcut/texture biases Tu et al. (2025); Xing et al. (2025); Geirhos et al. (2020; 2019)—all at test time, before any adaptation. The closest robustness findings under fine-tuning, LP-FT Kumar et al. (2022) and WiSE-FT Wortsman et al. (2022), show that clean-trained models lose out-of-distribution accuracy when they meet unseen shifts, and repair the loss by head-first staged training or by interpolating zero-shot and fine-tuned weights. What unites both threads is that robustness is an aggregate—a single top-1 number over the whole distribution. We differ on three axes. First, the shift (naturally occurring image blur) is present in the training mixture, yet adapting on degraded data does not by itself preserve competence on that degradation—it can selectively destroy it, an outcome invisible to corruption-robust-training intuition. Second, the failure is stratified, visible only when accuracy is decomposed by degradation level. Third, its cause lies in the interaction between adaptation direction and per-stratum decision geometry rather than a distribution-wide penalty. Whether interpolation remedies transfer to in-distribution stratified failure remains open.

Fine-tuning dynamics and representation geometry. Analyses of fine-tuned transformers—layer-wise attention drift Nguyen et al. (2021), including Full-vs-LoRA drift on clean data Xia (2026), neural collapse Papyan et al. (2020), catastrophic forgetting Kirkpatrick et al. (2017), and dimensional collapse Jing et al. (2022)—describe how training reshapes representation geometry. None connects optimization strength and adaptation depth to degradation-stratified robustness; our contribution is precisely that connection. Our analysis (Sec. 5) links a drift-magnitude jump, effective-rank contraction, and stratum-differentiated CKA divergence to the selective failure, and rules out a single-fragile-direction account.

Test-time vs. training-time adaptation. Test-time methods such as CLIPTTA Lafon et al. (2025) and Uniformity First Adachi et al. (2026) recover robustness at inference, so they are complementary to—not competing with—our training-time depth$\times$LR question. Likewise, CLIP domain-adaptation methods such as PADCLIP Lai et al. (2023) optimize aggregate accuracy or transfer on standard fine-grained benchmarks Deng et al. (2009); Van Horn et al. (2018); Maji et al. (2013); Krause et al. (2013); we expose the depth-conditional, stratum-specific sensitivity those aggregate metrics conceal.

3 Discovery: The Depth–LR Landscape

3.1 Experimental Setup

Model. We use CLIP ViT-B/32 Radford et al. (2021) as the backbone, with OpenCLIP Cherti et al. (2023) pretrained weights.

Dataset. Our primary dataset is a 20-class fine-grained bird classification task derived from iNaturalist Van Horn et al. (2018). The dataset comprises naturally occurring image degradation in real field photographs (not synthetically generated), stratified by an automated blur score into *sharp* (no detectable blur), *slight blur* (minor degradation), and *moderate blur* (strong degradation). The test set contains 11,232 images. The training split contains both natural photos and 36.1% synthetically chained degradations (including 6,218 directional-motion chains), so degraded inputs are substantially present at training time; the failure we study is not an unseen-shift effect (composition details in Appendix A). Blur strata are detector-derived with no manual labels; Appendix B audits that the slight-blur gap is a real-content selective failure rather than a labeling artifact.

Adaptation protocols. We sweep adaptation depth $L \in \{\text{L1}, \text{L2}, \text{L4}, \text{L6}, \text{L8}, \text{Full}\}$ (Lk = last k blocks unfrozen; Full = all blocks) and learning rate $\eta \in \{5\times10^{-6}, 10^{-5}, 2\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$. Writing a configuration as $c = (L, \eta)$, we define the stratum-d accuracy and selective degradation



Figure 2: **Depth–LR landscape** $(L,\eta)\to A_d\to G_{\text{sel}}$ **(Bird, seed 42).** Red marks the high-risk corner Full@10^{-4} ($G_{\text{sel}}=+19.5$pp).

gap:

$$A_d(L,\eta)=\frac{1}{|D_d|}\sum_{x\in D_d}\mathbf{1}[f_{L,\eta}(x)=y],\quad G_{\text{sel}}(L,\eta)=A_{\text{sharp}}(L,\eta)-A_{\text{slight}}(L,\eta),\tag{1}$$

where D_d is the test subset at blur level $d\in\{\text{sharp, slight, moderate}\}$; $G_{\text{sel}}>0$ marks selective loss (clean competence preserved, degraded lost).

Three-level framework. We trace the chain $(L,\eta)\to A_d\to G_{\text{sel}}\to D$, ending in per-sample drift $D(L,\eta)=\mathbb{E}_x[1-\cos(z_{L,\eta}(x),z_0(x))]$ (Sec. 5). All controlled runs share five epochs and a common data order, so depth and LR are the only varied factors; key configurations replicate over three seeds (mean±std), with independent 10/20-epoch dose checks (Sec. 6). Causal statements are confined to the fixed-boundary counterfactuals of Sec. 5.3.

3.2 The Depth–Optimization Risk Landscape

Figure 2 maps the full $(L,\eta)\to A_d\to G_{\text{sel}}$ chain across 6 depths × 5 learning rates on Bird. Three observations emerge, each sharpening the central claim that *adaptation dose*, not any single hyperparameter, governs robustness.

Observation 1: the preferred adaptation depth depends on η. At low η, full-network adaptation is optimal; as η rises, the preferred depth retreats to L8, then L6 (Fig. 2b): the preferred depth depends on η, not on the task alone; the L^* trajectory is domain-dependent (Sec. 4).

Observation 2: sensitivity grows with depth at high LR. At $\eta=10^{-4}$ the selective gap $G_{\text{sel}}=A_{\text{sharp}}-A_{\text{slight}}$ widens monotonically with depth—L4=−3.4pp, L6=−8.5pp, L8=−10.3pp, Full=−19.5pp—so deeper adaptation amplifies sensitivity, and increasingly so at higher LR.

Observation 3: the failure is selective, not global. At Full@10^{-4}, sharp accuracy holds at 98.3% (three-seed mean $98.5\pm0.2\%$) while slight-blur collapses to 78.8%—the signature we call *degradation-selective sensitivity*. Aggregate top-1 (86.2%) hides it as mild underperformance. The loss is non-monotone in blur magnitude, and magnitude-controlled paired degradations reproduce it only under directional motion, not objectively stronger isotropic blur (Sec. 6).

3.3 Multi-Seed Reproducibility

Replicating key configurations over seeds 1337 and 2024, Full@10^{-4} drops 5.7 ± 0.2pp while L6 drops only 0.6 ± 0.1pp (about 9× smaller), leaving Full 9.7–10.2pp below L6 throughout (Appendix Fig. 8a). The Full−L6 penalty is -10.0 ± 0.24pp and the sharp−slight gap 19.7 ± 0.28pp, with a maximal within-entry std of 0.34pp—an order of magnitude below the gap—which rules out unlucky initialization (per-seed values in Appendix D).

4 CROSS-DOMAIN VALIDATION

We ask which part of the Bird finding survives a change of domain, replicating a 4×3 depth–LR matrix on two fine-grained datasets with *synthetic* degradation—Stanford Aircraft (100 classes) Maji et al. (2013) and Stanford Cars (196 classes) Krause et al. (2013)—at two controlled blur severities and three seeds (42/1337/2024; mean±std).

4.1 VARIATION IN SELECTIVE SIGNATURE ACROSS DOMAINS

Cross-domain replication separates two claims of unequal generality: the *risk* claim—high-dose Full is a high-risk regime—replicates in all three domains, whereas the *selective* claim—sharp preserved while degraded collapses—is strongest on Bird:

- **Naturally occurring degradation (Bird)**: the strongest signature—sharp $98.5 \pm 0.2\%$ vs. slight $78.8 \pm 0.3\%$, a 19.7 ± 0.3pp gap stable across seeds.
- **Synthetic degradation (Aircraft, Cars)**: *attenuated*, near-uniform drops, but the high-LR Full-vs-partial penalty is equally pronounced (-9.1 to -10.7pp): the risk landscape replicates even when selectivity is weak.

Because these domains change class count, image statistics, and the operator library simultaneously with real-vs-synthetic provenance, the cross-domain comparison alone cannot pin the attenuation on provenance. The paired within-domain control (Sec. 6) untangles one axis: on Bird's own sharp photos, the exact operators reproduce the Full-specific gap under directional motion blur (25.1pp vs. 4.4pp for L6) but not under stronger isotropic blur, so selectivity requires directional-motion geometry.

Geometry alone is insufficient: the observed selectivity also appears to require a suitable baseline condition. Bird's inverted baseline (sharp$-$slight $= -10.33$pp) leaves a near-boundary slight subpopulation to flip (m_0 -0.15 vs. sharp $+0.06$; Sec. 5.3), whereas Aircraft's near-floor zero-shot (16.98%, $+1.92$pp) and Cars' already-adjudicated slight ($+15.16$pp) start from the opposite state, tracked by the observed selectivity (S_{slight} $+30.0$/$+1.0$/-10.1pp); saturated Imagenette is near-zero. Within the tested domains, selectivity appears only where the baseline leaves substantial room for stratum-specific boundary crossing (Appendix E.1). Risk generality, in short, runs broader than selectivity generality.

4.2 CROSS-DOMAIN SUMMARY

Table 1 reports the full adaptation-depth × learning-rate landscape across all three domains and all three seeds.

What replicates. Two properties hold everywhere. First, preferred depth is optimization-dependent rather than a fixed architectural choice: it holds at L8 or retreats shallower (L8→L6 on Bird) as LR rises, and never does deeper adaptation win at the highest LR. Second, the Full–BestPartial penalty grows monotonically with LR in every domain, from -0.9–-2.2pp to -9.1–-10.7pp (seed std ≤ 1.0pp; Table 1; Appendix Fig. 8b–c).

What does not replicate. The fine structure is domain-specific: the L8→L6 shift and the sharp–slight amplification are both Bird-specific (L8 stays best on Aircraft/Cars at every LR). The risk landscape itself survives everywhere.

Composition modulates, but does not cause, selectivity. Varying the Bird training composition at $\eta = 10^{-4}$ over three matched conditions (natural-full 52,968 rows, ∼67% slight; natural-sub 10,586; balanced 1+2+2 synthetic 10,585; three seeds), selectivity S_{slight} is $+30.0$/$+24.7$/$+61.2$pp (Full) vs. $+19.2$/$+15.5$/$+51.2$pp (L6; Table 23). Balanced composition *amplifies* selectivity, while halving the sample only mildly lowers it, so composition is a modulator, not a cause. Balanced Full *negatively* transfers on slight ($G = -5.8$pp) against $+40.4$pp under natural composition: a slight-majority set forces learning slight, a slight-minority set frees sacrificing it. The selectivity *base* is thus fine-grained content under high-LR Full; that balanced synthetic composition *amplifies* selectivity on Bird yet Aircraft and Cars show none marks the baseline state, not degradation geometry, as the necessary condition (Appendix E.1).

Table 1: **Adaptation-depth × learning-rate landscape across three domains and three seeds.** Slight-blur accuracy (%); mean±std over seeds 42/1337/2024 with individual values ([†]=Bird seed-42 only). Full@10^{-4}'s selective collapse on Bird (slight ≈ 78.8%) reproduces as a high-LR penalty across Aircraft and Cars; L6/L8 stay robust.

		Bird				Aircraft				Cars			
Depth	η	mean±std	S42	S1337	S2024	mean±std	S42	S1337	S2024	mean±std	S42	S1337	S2024
Full	2×10^{-5}	87.54[†]	87.54			75.28±0.59	74.74	75.19	75.91	83.31±0.13	83.17	83.40	83.37
	5×10^{-5}	84.54±0.14	84.52	84.69	84.42	73.47±0.53	74.05	73.36	73.00	80.12±0.35	79.75	80.18	80.44
	1×10^{-4}	78.81±0.34	78.79	79.16	78.49	68.08±0.18	68.17	67.87	68.20	73.22±0.43	72.73	73.52	73.41
L8	2×10^{-5}	89.47±0.26	89.70	89.19	89.53	76.79±0.39	77.20	76.75	76.42	84.26±0.36	84.43	83.85	84.50
	5×10^{-5}	88.99[†]	88.99			77.58±0.25	77.38	77.50	77.86	85.39±0.13	85.50	85.24	85.42
	1×10^{-4}	87.43±0.16	87.38	87.61	87.31	77.14±0.61	77.83	76.93	76.66	83.94±0.33	84.28	83.91	83.63
L6	2×10^{-5}	88.63[†]	88.63			73.25±0.23	73.51	73.15	73.09	81.81±0.33	82.13	81.48	81.83
	5×10^{-5}	89.43±0.15	89.59	89.40	89.29	75.59±0.45	75.70	75.97	75.10	83.68±0.20	83.71	83.86	83.47
	1×10^{-4}	88.79±0.25	88.98	88.88	88.50	74.86±0.63	75.31	75.13	74.14	82.89±0.21	83.10	82.89	82.69
L4	2×10^{-5}	83.94[†]	83.94			66.34±0.74	65.71	67.15	66.16	78.18±0.22	78.35	78.26	77.94
	5×10^{-5}	87.30[†]	87.30			71.32±0.38	70.90	71.41	71.65	81.80±0.30	81.81	82.09	81.49
	1×10^{-4}	87.55[†]	87.55			71.14±0.29	71.44	71.11	70.87	81.42±0.40	81.11	81.88	81.28

[†]Bird Full@2×10^{-5} was run only under seed 42 (n=1). All other entries use seeds 42/1337/2024.

4.3 Cross-Task Replication: A Non-Fine-Grained Domain

The cross-domain runs remain within fine-grained classification. To test whether the risk is an artifact of fine-grained label structure, we replicate on Imagenette—a 10-class coarse subset of ImageNet (9,469 train / 3,925 test)—with the identical synthetic-degradation protocol (one sharp, two slight and two moderate views per image; four operators balanced; seeds 42/1337/2024; Appendix Fig. 9).

Risk replicates, selectivity does not. On slight blur, Full drops to 94.85 ± 0.14%—below zero-shot slight (98.8%)—at η=10^{-4} while L6 holds 99.01 ± 0.17%, a Full–L6 gap of -4.16 ± 0.14pp stable across seeds: the depth–LR penalty is task-general. The sharp–slight gap on Full, however, grows only from 0.24 to 0.57pp against a zero-shot gap of 0.29pp ($\sim$ 0.3pp stratum-specific excess vs. 19.7pp on Bird), so selectivity is the fine-grained amplification while the depth–LR risk is the general component. Imagenette is zero-shot saturated (sharp 99.1%, slight 98.8%, moderate 95.4%; no headroom), so the 4pp Full–L6 divergence is pure damage. A non-saturated ImageNet-100 subset (zero-shot 68.5%; Appendix E.3) confirms: Full@10^{-4} regresses 8.4pp below zero-shot on slight blur while L6 gains +7.7–+25.2pp—risk generalizes, selectivity does not.

5 Representation-Level Analysis

To understand *why* the failure is selective, we analyze representations on a fixed 600-image subset (200 per blur level, seed 42); the phenomenon is established on the full test sets (Sec. 3–4). Define per-sample drift and its scalar magnitude

$$\Delta z = z_{\text{ft}} - z_{\text{pre}}, \quad D_{\text{rep}} = 1 - \cos(z_{\text{ft}}, z_{\text{pre}}), \tag{2}$$

with $z_{\text{ft}}, z_{\text{pre}}$ L2-normalized final-layer features. We report *primary* (is drift itself selective?), *supporting* (does Full additionally shift the geometry partials avoid?), and *auxiliary* evidence ruling out uniform confounders.

5.1 Primary Evidence: Drift Shows an Abrupt Jump at Full

On slight-blur features, partial depths at $\eta = 10^{-4}$ drift mildly and nearly depth-independently (D_{rep}: L4 0.708, L6 0.717, L8 0.718), whereas Full jumps to 0.799 (5×10^{-5}) and 0.798 (10^{-4}), ≈29% beyond the partial level, on all strata (Fig. 3a; full values in Appendix F). Cosine drift saturates already at 5×10^{-5}; the further accuracy loss at 10^{-4} is accompanied instead by stronger subspace contraction (below).

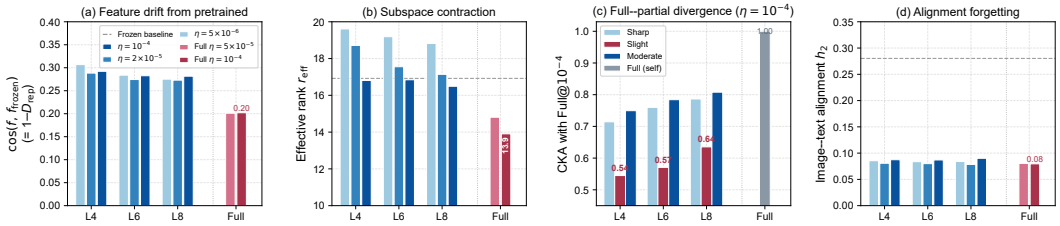


Figure 3: **Representation-level analysis** (slight-blur subset; same fixed sample). Full jumps drift D_{rep} above partials, contracts effective rank below the frozen baseline, and diverges from partials most on the slight stratum.

Linear CKA between adapted models shows partials remain mutually similar on slight blur (0.77–0.83), while Full diverges from every partial far more on slight blur (0.54–0.64) than on sharp inputs (0.71–0.79): the change is largest on the stratum whose accuracy collapses.

5.2 Supporting Evidence: Full Contracts the Subspace

Effective rank remains near the frozen baseline for partial adaptation (16.5–16.8 vs. frozen 16.9) but contracts under Full (14.8 at 5×10^{-5}, 13.9 at 10^{-4}), increasingly with dose (Fig. 3b). Absolute frozen rank varies across 200-sample subsets, so we rely on within-sample ordering: Full is lowest in every joint measurement (full values in Appendix F).

Summary and negative result. Across the three measurements, failing Full is geometrically distinct from every stable partial (associated with, not causal for, the collapse). Image–text alignment forgetting is likewise non-selective (Appendix F.1). A single fragile direction is also ruled out: across eight configs PC1 of centered drift explains only 9.3–11.1%, slight samples project *less* onto it than sharp (ratio 0.69–0.77), and $\|\Delta z\|$ is stratum-uniform (0.98–1.00). Selective failure thus stems from a distributed, stratum-blind change.

5.3 A Margin–Drift Mechanism

Matched drift magnitude (Sec. 5: ratio 0.98–1.00) rules out a generic magnitude-driven distortion and directs the explanation to drift *direction* against per-stratum decision boundaries. On the same 600 images we fit a frozen linear probe $h(z) = Wz + b$ (50.3% top-1) as a *controlled fixed boundary* for counterfactuals. Replaying drift along $z_\alpha = z_0 + \alpha\Delta z$, each pairwise gap is affine, $m_{yk}(\alpha) = m_{yk}(0) + \alpha q_{yk}$ with $q_{yk} = (w_y - w_k)^\top \Delta z$.

Adaptation tolerance (characterization). *For an initially correct sample under the fixed boundary, the first crossing occurs at*

$$\tau(x) = \alpha^*(x) = \min_{k:\, q_{yk}<0} \frac{m_{yk}(0)}{-q_{yk}}, \quad \text{flipped by } \alpha{=}1 \iff \min_k q_{yk} < -m_0(x). \tag{3}$$

We call τ the sample's *adaptation tolerance* (margin buffer per unit adverse drift); robustness is better characterized by τ than by $\|\Delta z\|$ within this fixed-boundary replay.

Margin buffer and its negative tail. Despite matched drift magnitude, slight-blur inputs start with lower frozen-probe margin (median m_0 -0.15 vs. $+0.06$; Fig. 4a): replaying Full@10^{-4} flips 26.6% vs. 8.7% of initially correct slight/sharp samples. The asymmetry survives margin matching—within $m_0 \in (1,2]$, 21.1% of slight vs. 0% of sharp carry adverse drift beyond their buffer—so the residual tail is high-dose-Full-specific (Appendix G).

Directional interventions (no retraining). Norm matching preserves a 3.8× slight/sharp flip ratio, ruling out magnitude. Random-direction controls show the learned direction is globally protective but less so for degraded inputs (69pp vs. 44pp), so failure is *insufficient protection*; τ transfers to the independently co-adapted head (AUC 0.96 vs. 0.76 single-normal; Appendix H). Training-distribution controls likewise exclude unseen-stratum and sample-size confounds (Appendix I, Table 17).

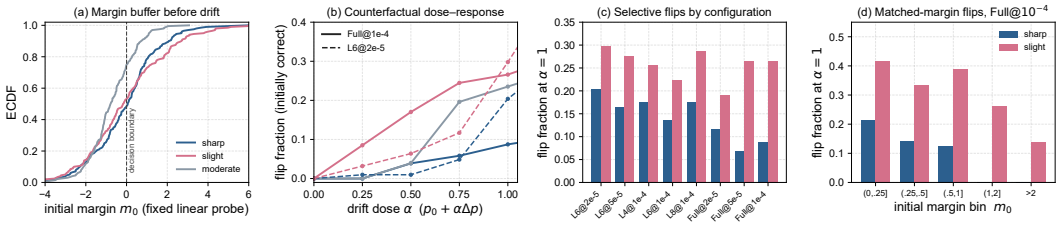


Figure 4: **Margin–drift mechanism on the fixed linear probe.** Low initial buffers **(a)**, dose-ordered flips **(b)**, high-LR-Full-specific selectivity **(c)**, and persistent matched-margin asymmetry **(d)** (measured, not conceptual).

As a boundary-level diagnostic, τ (Eq. 3) does not by itself account for why co-adaptive training leaves a persistent core—a question of *dose*, isolated next (Sec. 6).

6 Dose Dynamics and Content Control

We extend the three key configs to planned budgets $E \in \{5, 10, 20\}$ as *independent runs* (separate cosine schedules), emphasizing patterns monotone across budgets (Appendix J).

Movement continues after rescue saturates. For Full@10^{-4}, slight-blur accuracy plateaus after E=5 (78.8/80.4/80.3%) while representation drift keeps growing ($\|\Delta z\|$ 1.280/1.296/1.314); L6 shows the inverse pattern, with drift saturating (1.194/1.194/1.191) while slight-blur accuracy continues to recover (89.0/89.4/89.9%; Table 18). Mean drift directions stay aligned with E=5 ($|\cos|$>0.986), but Full@10^{-4}'s centered first principal direction rotates substantially by E=20 ($|\cos|$=0.32 vs. 0.90 for L6): extra Full dose redirects the movement plane rather than scaling it, so it stops rescuing slight errors.

Confidence rises without error recovery. Replayed on the *same* fixed boundary, adverse-projection fractions (0.18/0.13/0.14) and realized flips (26.6/25.5/24.5%) stay flat across dose while norm-matched random directions flip increasingly (protection gap 44 → 55pp): tolerance is not exhausted. The co-adapted head's own slight margin nonetheless grows (5.05 → 9.54) as error plateaus (Table 19).

The persistent core is set by response direction, not initial margin. Sorting slight samples (n=200/config) by own-head correctness across $E \in \{5, 10, 20\}$ yields *persistent* (wrong at all three: 17/22/14), *rescued* (wrong only at E=5: 14/9/8), and *stable* groups. The groups are indistinguishable at start—frozen-probe margin −0.92 vs. −1.39 (overlapping CIs), zero-shot −0.015 vs. −0.016, blur score 0.600—but diverge in later logit direction (E=5 → 20): persistent errors gain +0.18 on the true class but +2.49 on the competitor (margin −2.35 → −4.53), whereas rescued errors gain +5.12 vs. −0.20 (Fig. 5). Persistence is non-monotone in severity (0% sharp, 0.5–1.0% moderate, 7–11% slight). Depth changes the failure mode but neither rescues (Appendix J). Thus, the remaining error core is associated with late response direction rather than initial margin.

Minimal existence model. A minimal two-dimensional model (Appendix J, Fig. 13) shows matched-margin persistent errors arise when late updates follow a bulk-supported direction that does not rescue them.

Content-controlled degradation on the sharp test photos. To disentangle provenance from class count, image statistics, and operator library, we apply the exact four Aircraft/Cars operators (directional motion blur, defocus, camera shake, JPEG compression; slight/moderate) to *all* 474 *sharp benchmark test photos* and re-evaluate the same ten models on paired variants (Table 22); only degradation geometry changes. Zero-shot CLIP drops 28.3 → 18.4% on slight variants, yet Full's loss is anisotropy-specific: at Full@10^{-4}, E=5, synthetic motion blur reproduces the selective failure (25.1pp [21.3,28.9]) while defocus, camera shake, and JPEG yield $|\Delta| \leq 0.2$pp (L6: 4.4pp motion). The effect tracks degradation geometry, not provenance (23.3pp on `inat` vs. 26.0pp others; Appendix J).

Blur magnitude alone does not explain the observed selectivity. A Laplacian-variance analysis shows synthetic slight defocus is *stronger* than the real slight-stratum photos yet costs Full 0.0pp,

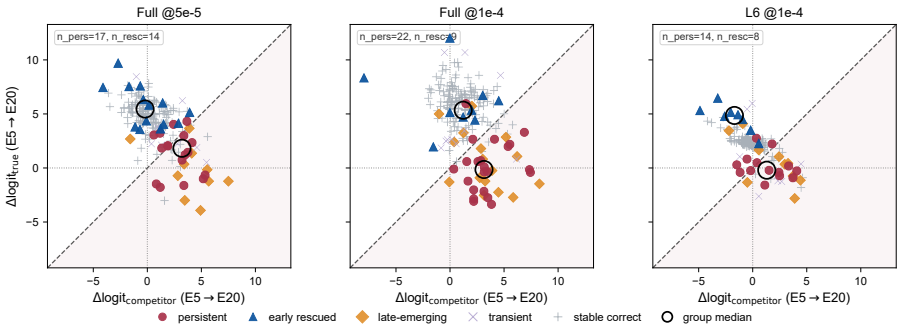


Figure 5: **Per-sample logit growth, $E{=}5 \to 20$, slight blur ($n{=}200$/config).** Persistent errors (circles) lie below the diagonal: competitor logits grow at least as much as true-class logits, so extra adaptation deepens rather than closes their gap. Rescued errors (triangles) sit upper-left; circles mark group medians.

while synthetic slight motion matches their strength and costs 25.1pp (Appendix J): selectivity tracks directional motion, not blur magnitude or provenance.

What remains specific to real content. The pooled severity inversion is a detector-label composition artifact (Appendix B); the real persistent core never recovers with dose, while the synthetic-motion gap partly closes; and balanced composition amplifies selectivity (Sec. 4), so natural composition buffers but does not cause it.

7 Mitigation Boundary and a Simple Selection Protocol

Two mitigation attempts fail. Two training-time interventions fail to recover a safe Full configuration: DADU's gates collapse to a layer-constant, degradation-independent profile (Appendix Fig. 14), and TDRC's alignment-preserving loss trades sharp for slight without reaching L6-level robustness (Appendix K).

A simple empirical selection protocol. Four validation-only rules score all 30 Bird configs (Appendix L); the three degradation-aware rules select L8@$2{\times}10^{-5}$ (top-1 93.0%, $G_{\text{sel}} = -7.6$pp vs. -19.5pp Full@10^{-4}), while sharp-only alone fails (-13.2pp). A pre-registered 50/50 replication on Aircraft/Cars confirms all four rules select L8@$5{\times}10^{-5}$ (top-1 77.5/86.7% vs. 68.1/75.0% for avoided Full@10^{-4}).

8 Discussion

Implications. Adaptation depth controls the gain–damage regime: at high LR Full crosses into net loss—on coarse tasks, below zero-shot—while partials retain gains, since replay τ (Eq. 3) better characterizes robustness than drift norm. The contribution is geometric understanding of adaptation, not a robust-fine-tuning recipe.

Limitations. **(1) Scope:** one backbone and blur degradations. **(2) Causality:** limited to replayed fixed-boundary interventions. **(3) Mechanism:** τ validated only on Bird. **(4) Content control:** the naturally occurring degraded composition is unreproducible from sharp photos; the selective *base* remains correlational (Sec. 4).

REPRODUCIBILITY STATEMENT

We describe all ingredients needed to reproduce our results. The datasets (Bird, FGVC-Aircraft, Stanford Cars), the four degradation operators, model configurations, and pre-processing are specified in Section 3 and Appendix A; training hyperparameters (adaptation depths, learning rates, seeds 42/1337/2024, epoch budgets, data order) are listed in the same appendix. Every main-text number is backed by per-configuration evaluation tables (Appendix D–L), and the fixed-boundary, margin-intervention, dose, persistent-core, and paired degradation analyses are accompanied by per-sample CSVs documenting each stratum. All training/evaluation scripts, configuration files, extracted-feature caches, figure-generation code, and model checkpoints will be released in an anonymized repository upon acceptance, enabling end-to-end reproduction of every figure and table.

AI USE STATEMENT

We used AI-assisted writing tools (large language models) solely to aid and polish the language of this manuscript. These tools were not used to generate research ideas, design experiments, produce or analyze results, or create references. All scientific content, empirical findings, and conclusions are the work of the authors, who verified all AI-generated text and take full responsibility for the content of this paper.

REFERENCES

Kazuki Adachi, Shin'ya Yamaguchi, and Tomoki Hamagami. Uniformity first: Uniformity-aware test-time adaptation of vision-language models against image corruption. *Transactions on Machine Learning Research*, 2026.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

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

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

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*, 2022.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3 D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Marc Lafon, Gustavo Adolfo Vargas Hakim, Clément Rambour, Christian Desrosier, and Nicolas Thome. ClipTTA: Robust contrastive vision-language test-time adaptation. In *Advances in Neural Information Processing Systems*, 2025.

Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. PADCLIP: Pseudo-labeling with adaptive debiasing in CLIP for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16109–16119, 2023.

Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 14394–14410, 2024.

Mushui Liu, Bozheng Li, Jun Dan, Ziqian Lu, Zhao Wang, and Yunlong Yu. Fully fine-tuned CLIP models are efficient few-shot learners. *Knowledge-Based Systems*, 324:113819, 2025.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013.

Ujjwal Mishra, Vinita Shukla, Praful Hambarde, and Amit Shukla. Improvise, adapt, overcome – telescopic adapters for efficient fine-tuning of vision language models in medical imaging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7605–7615, 2026.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? Uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Greg Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 8748–8763, 2021.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 43685–43704, 2024.

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*, 2020.

Weijie Tu, Weijian Deng, and Tom Gedeon. Toward a holistic evaluation of robustness in CLIP models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:8280–8296, 2025.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Chenguang Wang, Ruoxi Jia, Xin Liu, and Dawn Song. Benchmarking zero-shot robustness of multimodal foundation models: A pilot study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ruize Xia. Matched-learning-rate analysis of attention drift and transfer retention in fine-tuned CLIP. *arXiv:2604.16410*, 2026.

Shuo Xing, Lanqing Guo, Hongyuan Hua, Seoyoung Lee, Peiran Li, Yufei Wang, Zhangyang Wang, and Zhengzhong Tu. Demystifying the visual quality paradox in multimodal large language models. *arXiv:2506.15645*, 2025.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16795–16804, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

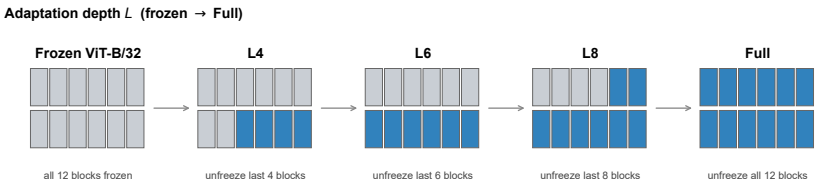


Figure 6: Adaptation depth L: the number of unfrozen transformer blocks, ranging from frozen (prompt-tuning regime) to Full (all 12 blocks).

A Experimental Setup

Model. We use CLIP ViT-B/32 Radford et al. (2021) as the backbone, with OpenCLIP Cherti et al. (2023) pretrained weights downloaded from the official repository. The model has 12 transformer blocks and 150M parameters.

Datasets.

- *Bird*: A 20-class fine-grained bird classification task derived from iNaturalist Van Horn et al. (2018). Contains naturally occurring image degradation in *real* field photographs (not synthetically generated). Stratified by blur level via an automated sharpness detector (Laplacian/Tenengrad/Brenner/high-frequency energy; it measures apparent blur, not its physical cause): *sharp* (no detectable blur), *slight* (minor degradation), *moderate* (strong degradation). Test set: 11,232 images. The training split (52,967 images) contains both natural photos and 36.1% synthetically chained degradations (compression, defocus, motion, camera shake, mixed; 6,218 motion chains), so degraded inputs are present at training time—the failure studied in the main text is not an unseen-shift effect.
- *Stanford Aircraft* Maji et al. (2013): 100 classes, 3,333 test images. We apply controlled *synthetic* Gaussian blur at two severity levels ($\sigma = 1.0$ and $\sigma = 2.0$); each test image is evaluated under sharp and both blurred conditions (9,999 evaluated images).
- *Stanford Cars* Krause et al. (2013): 196 classes, 8,041 test images. Same synthetic blur protocol as Aircraft.

Adaptation protocols. We sweep two hyperparameters:

- *Adaptation depth* $d \in \{\text{L1}, \text{L2}, \text{L4}, \text{L6}, \text{L8}, \text{Full}\}$, where L$k$ means unfreezing the last k transformer blocks; Full means unfreezing all 12 blocks. The projection head (image encoder, text encoder) is always fine-tuned alongside the unfrozen blocks.
- *Learning rate* $\eta \in \{5\times10^{-6}, 10^{-5}, 2\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$.

Training details. All experiments use 5 training epochs, AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay 0.2), batch size 32, and cosine learning-rate schedule with 5% warmup. Data augmentation during training: random resize crop (224×224), random horizontal flip. All Bird experiments use seed 42 for the discovery sweep; key configurations are replicated with seeds 1337 and 2024. Cross-domain experiments (Aircraft, Cars) use all three seeds.

Evaluation. We evaluate on the full held-out test set and report Top-1 accuracy decomposed by degradation level. The *degradation gap* is defined as $\Delta_{\text{deg}} = A_{\text{slight}} - A_{\text{sharp}}$ (pp).

B Label Reliability Audit

The Bird blur strata are detector-derived—five normalized focus metrics (Laplacian variance, Tenengrad, Brenner, high-frequency ratio, autocorrelation) are averaged and thresholded at $0.75/0.50/0.25$—with no manual labels. Because the headline claim is a *slight*-specific gap, we audit whether that gap could be a labeling artifact rather than a genuine selective failure. All

Table 2: Pooled-vs-real split of the load-bearing strata (Bird, seed 42), Full@10^{-4}. The slight gap is real-content-driven; the moderate "inversion" is a composition artifact.

stratum	population	Full@10^{-4} acc (%)	Full−L6 gap (pp)
slight	pooled	78.8	−10.2
slight	real-iNat (n=4,450)	66.7	−16.7
moderate	pooled (1,207 chain + 46 real)	97.8	+0.3
moderate	real-iNat (n=46)	54.4	—

quantities are per-image predictions from the evaluation pipeline with 2,000-resample bootstrap CIs (seed 42), reported on the pooled test set and on the real iNaturalist photos alone. Source CSVs: `results/mechanism/label_audit_a*.csv`.

Threshold robustness (T1). Across 25 legal cutoffs (slight $\in [0.65, 0.85]$, moderate $\in [0.40, 0.60]$) the Full−L6 gap on slight blur is negative in every case (pooled −8.3 to −11.3pp; real −16.2 to −17.2pp), and removing samples within ±0.10 of a cutoff *increases* the effect (pooled −11.0, real −17.2pp). The gap is not a boundi-cutoff artifact.

Continuous-score profile (A3). In fixed-width 0.05 bins of the continuous blur score the gap forms a mid-range bulge concentrated at 0.55–0.70 (pooled −8.6 to −12.8pp), whereas bins immediately adjacent to the cutoffs (0.50–0.55, 0.70–0.75) show Full ≈ L6. A cutoff artifact would peak at the boundaries; the opposite holds.

Detector reproducibility (T2). The blur score is exactly reconstructed from the five raw metrics (MAE = 0.0000, n=8,813). Leaving any single metric out keeps the real-iNat gap at −16.3 to −17.9pp.

Ground-truth label error (T3). Of Full@10^{-4}'s 1,503 slight errors, 68.5% [66.1, 70.9] have at least two independent models predicting the annotated class—*higher* consensus support than L6's errors (38.5%), so Full's failure is "alone wrong," not "question wrong." In the disagreement set (Full wrong, L6 right, n=905), zero images lack any independent model supporting the label, and 95.9% have at least two. Explaining the gap by mislabeling would require 723 systematic mislabels, of which the candidate count is zero.

Level–type confounding (T4). Within slight blur the gap holds separately for camera-shake (−10.0pp) and the remaining type (−10.8pp); the level split does not proxy a single degradation type.

Real-vs-synthetic provenance split. The 8,813 non-synthetic rows comprise 4,655 real iNaturalist photos and 4,158 chained-synthetic rows (source tagged `unknown` but detector-scored). Full is near-perfect on chained-synthetic slight (99.2%, gap +0.8pp) and fails almost entirely on real photos (−16.7pp; Table 2): the slight gap is a real-content phenomenon and is *stronger* when measured on real content alone. The moderate stratum, by contrast, is composition-dominated (1,207 chained rows at 99.4% vs. 46 real rows at 54.4%), so the pooled "moderate easier than slight" inversion does not survive on real content.

Scope. The 68.5% consensus is model evidence, not a human gold standard; a small double-blind human audit of the disagreement set remains the only missing step. The real moderate stratum has n=46, so its entries carry wide CIs. The A3 logistic interaction (−0.36) is a linear approximation to the bulged relationship; the bin profile above, not the interaction coefficient, is the appropriate summary.

C Complete Depth×LR Results (Bird, Seed 42)

Figure 7 visualizes the complete depth×LR matrix for Bird at seed 42 across four metrics. Table 3 gives the exact numerical values.

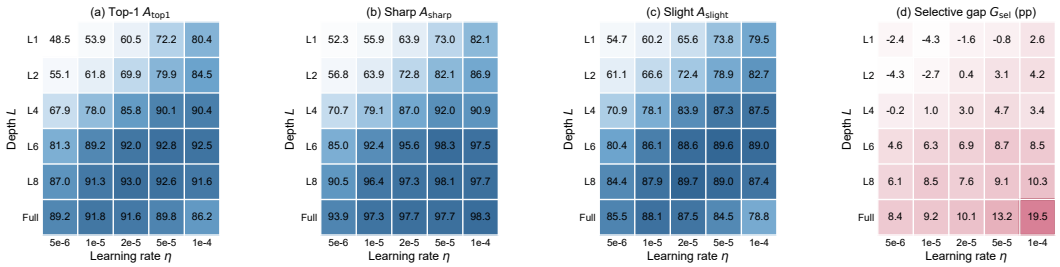


Figure 7: Complete Bird depth×LR landscape (seed 42). (a) Top-1 accuracy $A_{\text{top1}}(L,\eta)$. (b) Sharp accuracy A_{sharp}. (c) Slight-blur accuracy A_{slight}—the failing stratum collapses at Full@10^{-4}. (d) Selective gap $G_{\text{sel}} = A_{\text{sharp}} - A_{\text{slight}}$ (pp); red marks the high-risk corner. Red box = Full@10^{-4}.

Table 3: Bird seed 42, all depths × all LRs. *Left block*: slight-blur Top-1 accuracy (%). *Right block*: degradation gap $\Delta_{\text{deg}} = A_{\text{slight}} - A_{\text{sharp}}$ (pp); more negative = more degradation-sensitive.

	A_{slight} (%)					Δ_{deg} (pp)				
Depth	5e-6	1e-5	2e-5	5e-5	1e-4	5e-6	1e-5	2e-5	5e-5	1e-4
L1	54.7	60.2	65.6	73.8	79.5	+2.4	+4.3	+1.6	+0.8	−2.6
L2	61.1	66.6	72.4	78.9	82.7	+4.3	+2.7	−0.4	−3.1	−4.2
L4	70.9	78.1	83.9	87.3	87.6	+0.2	−1.0	−2.1	−4.7	−3.4
L6	80.4	86.1	88.6	**89.6**	**89.0**	−4.6	−6.3	−6.9	−8.7	−8.5
L8	84.4	87.9	**89.7**	89.0	87.4	−6.1	−8.5	−7.6	−9.1	−10.3
Full	**85.5**	**88.1**	87.5	84.5	78.8	−8.4	−9.2	−10.1	−13.2	**−19.5**

D Per-Seed Replication

The Full–BestPartial penalty grows monotonically with η across all three domains (Fig. 8b; 3 seeds, error bars ±1 std). Tables 4 and 5 below give the exact per-seed values behind these panels.

E Cross-Domain Depth×LR Matrices

Table 6 presents the complete depth×LR matrices for Aircraft and Cars (synthetic degradation), mean over three seeds.

E.1 Baseline Precondition for Selectivity

Table 7 reports, per domain, the zero-shot top-1, the zero-shot sharp−slight gap, and the injected selectivity $S_{\text{slight}} = (A^{\text{ft}}_{\text{sharp}} - A^{\text{ft}}_{\text{slight}}) - (A^{\text{zs}}_{\text{sharp}} - A^{\text{zs}}_{\text{slight}})$ at Full@10^{-4} (three seeds). The absence of selectivity on Aircraft/Cars is a consequence of these baselines, not an unexplained deficit. Bird's selectivity emerges atop an *inverted* baseline—slight is already easier than sharp zero-shot (−10.33pp)—leaving a near-boundary-but-correct slight subpopulation whose frozen-probe margin is already low (median m_0 −0.15 vs. sharp +0.06; Sec. 5.3), so adaptation drift can flip it. Aircraft and Cars start from the opposite state: Aircraft's zero-shot is far below usable (16.98% on 100 classes) with sharp and slight nearly indistinguishable (+1.92pp), offering no boundary structure to flip, while Cars' zero-shot has already adjudicated slight (+15.16pp), placing its slight subpopulation on the wrong side of the boundary. Consistently, Full@10^{-4} injects near-zero net selectivity on both ($S_{\text{slight}} = +1.00$pp Aircraft, −10.10pp Cars vs. +30.04pp Bird), the negative Cars value meaning Full *closes* the pre-existing gap. Saturated Imagenette conforms (+0.29 → +0.28pp; Sec. E.2). Two bounds qualify this reading: the near-boundary account of Aircraft/Cars extends the margin–drift mechanism *by extension*, m_0 being measured on Bird alone and their subpopulation structure inferred from zero-shot accuracy; and the inverted Bird baseline is a base fact whose origin—plausibly the naturally degraded composition of iNat—is not explained here.

Table 4: Per-seed results (Bird). Slight-blur Top-1 accuracy (%).

Configuration	Seed 42	Seed 1337	Seed 2024	Mean±std
Full@5e-5	84.52	84.69	84.42	84.54±0.14
Full@1e-4	78.79	79.16	78.49	78.81±0.34
L6@5e-5	89.59	89.40	89.29	89.43±0.15
L6@1e-4	88.98	88.88	88.50	88.79±0.25
L8@2e-5	89.70	89.19	89.53	89.47±0.26
L8@1e-4	87.38	87.61	87.31	87.43±0.16
Full−L6 @1e-4	−10.19	−9.72	−10.01	**−10.0±0.24**
FullΔ_{HLR} (5e-5→ 1e-4)	5.73	5.53	5.93	5.73±0.20
Full sharp−slight @1e-4	19.5	19.6	20.0	19.7±0.28

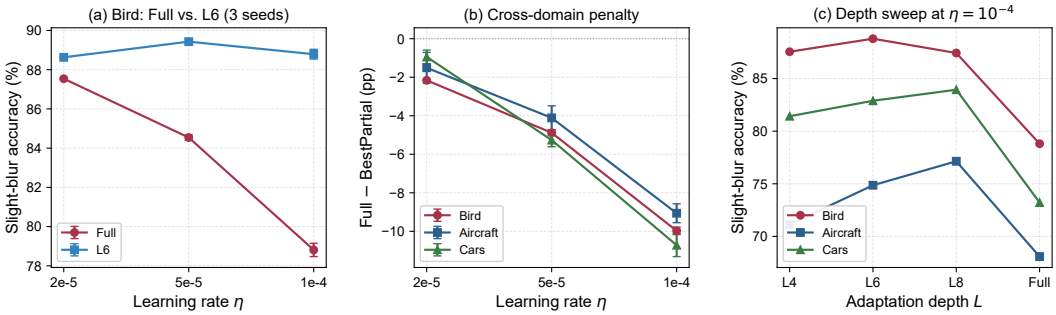


Figure 8: **Reproducibility and cross-domain generalization** (moved from the main text). (a) Bird Full vs. L6 slight accuracy across three seeds. (b) Full–BestPartial penalty grows monotonically with η in all three domains (-9.1 to -10.7pp at high LR). (c) Depth sweep at $\eta{=}10^{-4}$. The former panel (d) on DADU gate collapse is covered by Fig. 14 and Table 24; all exact values are in Tables 1, 4, and 5.

E.2 Cross-Task Replication (Imagenette)

Figure 9 reports the cross-task replication of Section 4 on Imagenette—a 10-class coarse subset of ImageNet (9,469 train / 3,925 test)—using the identical synthetic-degradation protocol (one sharp, two slight and two moderate views per image; four operators balanced; seeds 42/1337/2024). The depth–LR damage persists (-4.16 ± 0.14pp at $\eta{=}10^{-4}$, seed-stable) while the sharp–slight selective gap collapses below 0.6pp, confirming the main-text decomposition: the risk is task-general, selectivity is the fine-grained-domain amplification.

E.3 Cross-Task Replication: ImageNet-100 (Non-Saturated Gain-Side)

Imagenette (Sec. 4, Appendix E.2) is zero-shot saturated (sharp 99.1%, moderate 95.4%), so it cannot resolve whether the absence of selectivity reflects a true task property or ceiling compression. We therefore run the identical protocol on ImageNet-100—a 100-class coarse subset of ImageNet with substantial headroom (zero-shot top-1 68.46%). Training uses 300 images per class with one sharp, two slight and two moderate degradations per image (150,000 rows); evaluation uses the full 5,000-image validation set (25,000 rows). Table 8 reports the Full/L6 $\times$ $\{5{\times}10^{-5}, 10^{-4}\}$ matrix (one seed; the depth–LR gaps are $\geq$ 12pp, so direction is robust to seed noise).

Three conclusions follow. **(1) Risk generalizes.** Full–L6 at $\eta{=}10^{-4}$ ranges from -12.1pp (moderate) to -17.8pp (sharp), reproducing the depth–LR penalty on a non-fine-grained task with ample gain headroom. **(2) Selectivity is conditional, not universal.** The sharp–slight gap stays between 1.9 and 3.6pp across all four configurations, below the zero-shot baseline of 7.4pp; fine-tuning compresses the pre-existing asymmetry rather than amplifying it. No Bird-style selective collapse appears. **(3) Depth controls the gain–damage regime.** Doubling the learning rate costs L6 only $\sim$2pp uniformly, but costs Full 7.5–9.6pp; L6 stays strictly above zero-shot on every stratum at both rates ($+7.7$ to $+25.2$pp), whereas Full flips from small gains at $5{\times}10^{-5}$ to -8.4pp on slight

Table 5: Cross-domain per-seed results (mean±std over 3 seeds, pp).

Metric	Bird	Aircraft	Cars
Full−BestPartial @ 2e-5	$-2.2^{\dagger}$	−1.5±1.0	−0.9±0.4
Full−BestPartial @ 5e-5	−4.9±0.2	−4.1±0.8	−5.3±0.4
Full−BestPartial @ 1e-4	$\mathbf{-10.0 \pm 0.2}$	$\mathbf{-9.1 \pm 0.6}$	$\mathbf{-10.7 \pm 0.7}$
Δ_{HLR} (2e-5→ 1e-4)	$8.8^{\dagger}$	7.2±0.6	10.1±0.3
Full−L6 @ 1e-4	−10.0±0.2	−6.8±0.7	−9.7±0.6
d^* trajectory	L8→L6	L8 (stable)	L8 (stable)

†Bird Full@2e-5 was run only under seed 42 (n=1). All other entries use seeds 42/1337/2024.

Table 6: Aircraft and Cars (synthetic blur, 3-seed mean). Slight-blur Top-1 accuracy (%).

	Aircraft			Cars		
Depth	2e-5	5e-5	1e-4	2e-5	5e-5	1e-4
L4	66.3	71.3	71.1	78.2	81.8	81.4
L6	73.3	75.6	74.9	81.8	83.7	82.9
L8	**76.8**	**77.6**	**77.1**	**84.3**	**85.4**	**83.9**
Full	75.3	73.5	68.1	83.3	80.1	73.2

at 10^{-4}. Together with the bird results, this separates *adaptation risk* (broadly induced by excessive adaptation) from *degradation-selective failure* (conditional on task and degradation geometry).

F Representation Analysis

Table 9 reports the complete representation analysis across depths and LRs on the slight-blur test subset. Rows at η=10^{-4} and the two Full rows share a single fixed 600-image extraction (200 per blur level), so frozen, partial, and Full values are directly comparable; lower-LR partial rows are from the earlier partial-only extraction.

Table 10 reports linear CKA between adapted models within the joint sample. Full diverges from every partial depth most strongly on slight blur (the failing stratum), less on sharp inputs.

Table 11 reports the directionality test of Section 5: for each configuration, v_1 is the first principal component of the centered per-sample drift $\Delta f = f_{\text{adapted}} - f_{\text{frozen}}$ on the fixed 600-image test subset (200 per blur level, seed 42), and projection energies are computed on adapted features. The pooled Spearman correlation between projection energy and accuracy across the 24 (configuration, stratum) pairs is +0.24; it is positive because Full configurations drift more overall while sharp accuracy remains high, and it disappears under within-configuration stratum comparisons.

Figure 10 visualizes these quantities; together they falsify the single-fragile-direction hypothesis. Layer-wise drift analysis and per-degradation measurements are available in the project repository.

F.1 Alignment Forgetting Is Non-Selective

Figure 11 shows image–text alignment (h_2, cosine similarity between image and text embeddings) across adaptation depths at $\eta = 10^{-4}$ (joint same-sample run). Alignment forgets sharply already at L4 (0.280 → 0.087) and is only mildly lower for Full (0.080), so it cannot by itself discriminate the selectively failing configuration—it is an auxiliary rather than discriminative mechanism.

G Margin–Drift Mechanism: Derivation and Full Results

This section gives the protocol, the exact derivation summarized in Section 5.3, and the complete per-configuration measurements. Everything reported here is measured on the same 600-image subset as the directional test (200 images per stratum, `test.csv`, seed 42), for the same eight configurations.

Table 7: **Zero-shot baselines and injected selectivity per domain (Full@10^{-4}, three seeds).** S_{slight} is the zero-shot-anchored selective gap; negative means the pre-existing gap closes.

Domain	zs top-1 (%)	zs sharp−slight (pp)	S_{slight} (pp)
Bird (natural)	28.67	−10.33	$+30.0 \pm 0.3$
Aircraft (synthetic)	16.98	+1.92	$+1.0 \pm 0.1$
Cars (synthetic)	49.32	+15.16	-10.1 ± 0.2
Imagenette (synthetic)	97.50	+0.29	$+0.28 \pm 0.03$

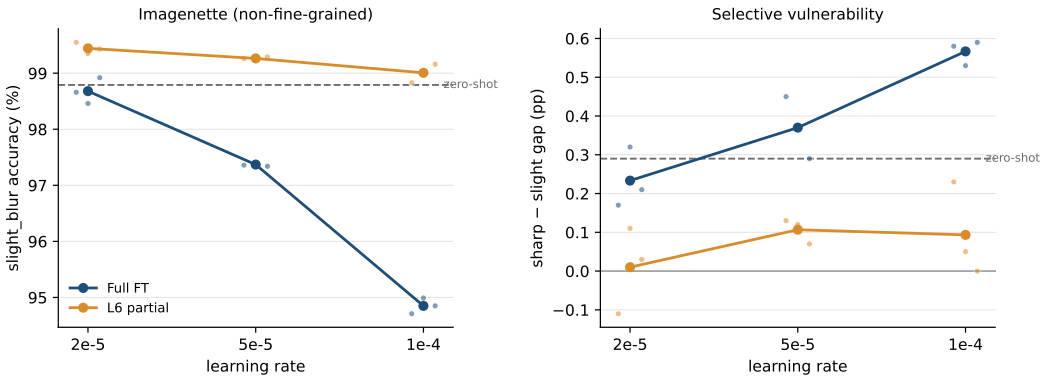


Figure 9: **Cross-task replication on Imagenette (non-fine-grained).** *Left*: slight-blur accuracy vs learning rate for Full FT and L6 partial (three seeds; zero-shot dashed). *Right*: sharp–slight selective gap vs learning rate. The depth–LR damage persists (-4.2pp at $\eta{=}10^{-4}$) while selectivity collapses below 0.6pp—risk is task-general, selectivity is fine-grained-specific.

Protocol and decision boundaries. Let $z_0(x)$ be the unnormalized pooled 512-d pretrained feature and $\Delta z(x) = z_{\text{ft}}(x) - z_0(x)$ each model's measured displacement. The classifier head is trained jointly with the encoder, so applying a deployed head to frozen features yields near-chance accuracy (6–10%): forward crossing along z_α is not defined for that head at $\alpha{=}0$. We therefore use three boundaries:

- **lp (primary):** a linear probe trained on frozen features (loaded from `checkpoints/linear.pt`), a competent fixed linear boundary with subset top-1 51.5/47.0/25.5% (sharp/slight/moderate) at $\alpha{=}0$. Along the interpolation $z_\alpha = z_0 + \alpha\Delta z$ its per-competitor gaps are exactly affine, so crossing doses are computed in closed form with no approximation.
- **zs (secondary):** the frozen zero-shot text-prototype cosine boundary (subset top-1 29/33/11%).
- **ft (endpoint only):** each configuration's own jointly trained head. At $\alpha{=}1$ its subset accuracies are 99.5/79.5/98.5%, exactly matching the deployed-model evaluation, which validates the feature pipeline; $\alpha{=}0$ is undefined as a fixed-boundary problem, so we use this head only for endpoint margins.

Exact derivation. For the fixed linear rule $h(z) = Wz + b$, write $g_k = w_y - w_k$ and $c_k = b_y - b_k$. The gap to competitor k along the path is $a_k(\alpha) = g_k^\top z_0 + c_k + \alpha\, g_k^\top \Delta z$. For an initially correct sample ($a_k(0) > 0\ \forall k$), each competitor crosses at $t_k = a_k(0)/(-g_k^\top \Delta z)$ whenever $g_k^\top \Delta z < 0$, hence

$$\alpha^*(x) = \min_{k:\, g_k^\top \Delta z < 0} \frac{g_k^\top z_0 + c_k}{-g_k^\top \Delta z}, \qquad \text{flip by } \alpha{=}1 \iff \min_k g_k^\top \Delta z < -m_0(x), \tag{4}$$

where $m_0(x) = \min_k a_k(0)$. The runner-up may change along the path; we therefore use the exact minimum, and report the first-order score $V(x) = -(\nabla m_0)^\top \Delta z / m_0(x)$ only as a per-sample predictor. The first-order margin change tracks the exact one with corr = 0.66–0.85 across configurations and strata (0.52 including the moderate-blur stratum at Full@10^{-4}), so the local-linear premise is approximate but serviceable. Optimization dose enters through $\Delta z \approx J_d \Delta\theta$ with $\|\Delta z\| \lesssim \eta \|J_d\| \|g\|$: η scales the displacement and depth enlarges the parameter paths through which it propagates.

Table 8: **ImageNet-100** 2×2 **(CLIP ViT-B/32, seed 42).** Zero-shot: sharp 84.2 / slight 76.8 / moderate 52.2 / top-1 68.5.

Stratum	ZS	Full@5e−5	Full@10^{-4}	L6@5e−5	L6@10^{-4}
sharp	84.2	80.0	70.4	90.1	88.2
slight	76.8	77.7	68.5	86.6	84.6
moderate	52.2	71.2	63.7	77.4	75.8
top-1	68.5	75.5	67.0	83.6	81.8

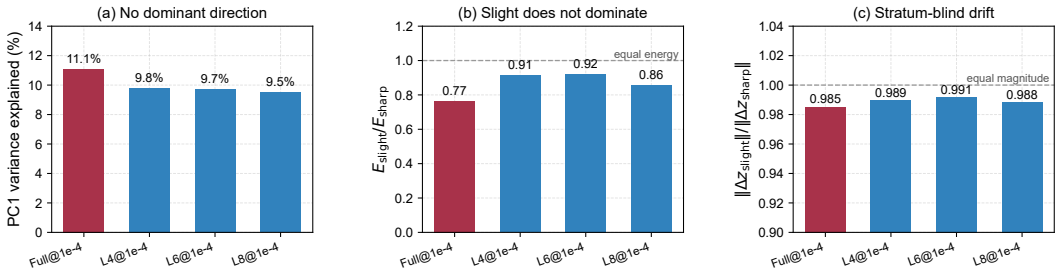


Figure 10: Directionality test (Bird, η=10^{-4}, slight-blur subset). (a) PC1 explains only ≈10% of centered drift variance—no single dominant direction. (b) Projection-energy ratio $E_{\text{slight}}/E_{\text{sharp}} < 1$: the failing stratum does not dominate drift energy. (c) Drift-magnitude ratio $\|\Delta z_{\text{slight}}\|/\|\Delta z_{\text{sharp}}\| \approx 1.0$: drift is stratum-blind. The falsification holds for Full and all partial depths.

Findings. (i) Drift is stratum-blind but margins are not: mean $\|\Delta z\|$ ranges 1.17–1.29 with slight/sharp ratio ≈ 0.99 in every configuration, while the slight-blur margin distribution sits left of the sharp one (medians -0.153 vs. $+0.061$; these m_0 are boundary-intrinsic and identical across configurations; Figure 4a). (ii) Selectivity is dose-ordered. At high LR Full flips 3.0–3.9× as many initially correct slight as sharp samples; every partial stays at 1.5–1.7×, and its crossings occur substantially later (median α^* among slight samples: 0.64–0.78 vs. 0.34–0.47 for high-LR Full). (iii) The first-order score predicts individual flips on slight blur (AUC 0.59–0.74, best at Full@10^{-4}) but is at chance on sharp (0.38–0.57): with large margin buffers, crossing is governed by margin size rather than the projected drift. (iv) Margin buffer alone is insufficient. Matching on $m_0 > 2$, Full@10^{-4} flips 13.9% of slight vs. 0% of sharp samples, while L6@10^{-4} flips 5.6% vs. 4.5%: an excess anti-margin drift component exists, and it is specific to the high-dose Full regime (Figure 4d). (v) The mechanism lives in the negative tail, not the mean. Against the fixed probe the mean margin *grows* along the path (m_1 medians exceed m_0 for both strata), and reversing the drift to α=-0.25 flips 54–100% of initially correct Full predictions (sharp/slight: 54–90%): the replayed drift improves separability on average while a larger slight-blur sub-population crosses. The deployed heads show the same ordering as endpoint margin compression (slight/sharp median ratio 0.66 at Full@10^{-4}, 0.70–0.82 at lower-LR Full, 0.87–0.94 for partials). (vi) Boundary dependence. No initially correct sample crosses the weak zero-shot boundary by α=1: both its margins and the cosine-space drift projections are an order of magnitude smaller, so the test there is uninformative rather than contradictory. We therefore claim a boundary-relative local mechanism, not a universal invariant.

Scope. The counterfactual replays measured, observational drift with a held-fixed classifier; it is not an intervention on training, and the linear-boundary analysis does not establish what happens along the actual head–encoder co-adaptation path beyond the endpoint margin ordering. The results nevertheless support a margin–drift interaction as a quantitative local explanation for how a distributed, stratum-blind representation change produces degradation-selective failure.

H Counterfactual Drift Interventions

This section decomposes the margin–drift mechanism with representation-space interventions on the same fixed linear probe and the same 600-image subset as Section G. No model is retrained: every counterfactual edits the measured displacement $\Delta z(x)$ and re-evaluates the frozen probe.

Table 9: Representation analysis (Bird, slight-blur). $\cos_{\text{frozen}}$: cosine similarity to frozen features (lower = more drift); r_{eff}: effective rank (participation ratio); h_2: image–text alignment. The 10^{-4} partial rows and both Full rows are from one joint same-sample run.

Depth	LR	Condition	$\cos_{\text{frozen}}$	r_{eff}	h_2
Frozen	—	slight-blur	1.000	16.92	0.280
L4	5×10^{-6}	slight-blur	0.307	19.61	0.086
L4	2×10^{-5}	slight-blur	0.288	18.71	0.081
L4	10^{-4}	slight-blur	0.292	16.81	0.087
L6	5×10^{-6}	slight-blur	0.284	19.19	0.084
L6	2×10^{-5}	slight-blur	0.274	17.55	0.080
L6	10^{-4}	slight-blur	0.283	16.85	0.087
L8	5×10^{-6}	slight-blur	0.275	18.82	0.084
L8	2×10^{-5}	slight-blur	0.273	17.14	0.078
L8	10^{-4}	slight-blur	0.282	16.49	0.090
Full	5×10^{-5}	slight-blur	0.201	14.81	0.080
Full	10^{-4}	slight-blur	0.202	13.90	0.080

Table 10: Linear CKA between adapted configurations (joint same-sample run).

Pair	sharp	slight-blur	moderate-blur
L4–L6	0.867	0.812	0.854
L6–L8	0.906	0.829	0.900
L4–L8	0.833	0.767	0.841
L4–Full	0.714	0.545	0.750
L6–Full	0.760	0.571	0.784
L8–Full	0.786	0.635	0.808
Full@5e-5–Full@1e-4	0.834	0.747	0.845

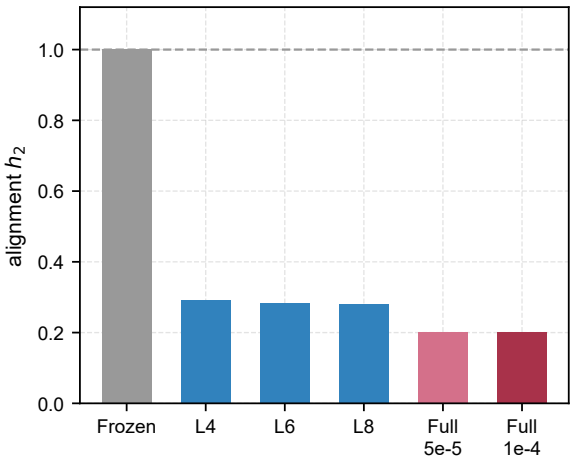
Construction. For each initially correct sample let $g(x) = w_y - w_{k^*}$ be the pairwise normal to its active competitor at α=0. We form (i) *actual*: $z_0 + \Delta z$; (ii) *parallel*: $z_0 + \Delta z_\parallel$ with $\Delta z_\parallel = (g^\top \Delta z/\|g\|^2)g$; (iii) *perpendicular*: $z_0 + \Delta z_\perp = (I - gg^\top/\|g\|^2)\Delta z$; (iv) *perp norm-matched*: the same $\Delta z_\perp$ rescaled per sample to $\|\Delta z\|$; (v) *norm-matched drift*: each Δz rescaled to the pooled median norm r=15.28 (unnormalized probe coordinates); and (vi) *random direction*: $z_0 + Q\Delta z$ for 20 fixed random orthogonal matrices (Gaussian QR, seed 42), which preserve norm exactly while destroying the learned directional structure. All flips below are conditioned on correctness at α=0.

Crossing CDF. With $\tau(x) = \inf\{\alpha : m_y(\alpha) \le 0\}$, the cumulative crossing probability is $P(\tau \le \alpha \mid d) = F_d(\alpha)$; this counts *at least one* crossing by dose α and equals current error only if crossings do not recover. On the grid $\alpha \in \{0.25, 0.5, 0.75, 1\}$ we record exactly zero cross-then-recover events among initially correct samples in all eight configurations, so on this grid the two coincide.

Findings. (i) *Magnitude is excluded.* Norm-matched drift retains the slight/sharp flip asymmetry at Full@10^{-4} (22.3% vs. 5.8%, 3.8×; Table 14 shows it is not Full-specific). (ii) *A single boundary normal is insufficient.* The $\Delta z_\parallel$ component along the initial runner-up normal accounts for only 11.7% of slight flips, while $\Delta z_\perp$ flips *more* samples than the actual drift (38.3% vs. 26.6%; Table 13): mean projection is protective (Section G finding v), and only 0–20% of realized crossings involve the initial runner-up. The multi-class piecewise-linear form $m_y(\alpha) = \min_{k \ne y}[m_{yk}(0) + \alpha q_{yk}]$ accounts for this—the active competitor switches along the path—which the binary single-normal model cannot express. (iii) *Margin-matched negative tail.* Within matched m_0 bins, slight-blur samples carry a heavier negative pairwise projection at Full@10^{-4} (Table 15); at L6@10^{-4} the same bins flip only 4.5–5.6% for either stratum—L6 survives because its smaller drift leaves these margins largely intact. (iv) *The learned direction is protective, asymmetrically.* Random norm-preserving directions flip 70–88% of predictions, an order of magnitude above the actual adaptation, so fine-tuning drift is not destructive in direction; it is strongly boundary-protective. Protection is however larger for sharp than slight inputs in every configuration (16–26pp paired gap; 25–26pp at high-LR Full), and sharp protection grows with adaptation dose (Full 0.64–0.70 vs. partials 0.43–0.50) while slight protection does not. (v) *Scope.* These results establish a causal role for directional

Table 11: Directionality test: v_1 variance explained, projection-energy ratios (slight/sharp, slight/moderate), and mean drift-magnitude ratio (slight/sharp).

Configuration	v_1 var. expl.	E_{sl}/E_{sh}	E_{sl}/E_{mod}	$\|\Delta f\|_{sl}/\|\Delta f\|_{sh}$
Full@2×10^{-5}	0.109	0.73	0.91	0.99
Full@5×10^{-5}	0.106	0.69	0.93	0.98
Full@10^{-4}	0.111	0.77	0.98	0.99
L4@10^{-4}	0.098	0.91	1.12	0.99
L6@10^{-4}	0.097	0.92	0.97	0.99
L8@10^{-4}	0.095	0.86	0.98	0.99
L6@2×10^{-5}	0.093	1.03	1.00	1.00
L6@5×10^{-5}	0.095	1.01	0.95	0.99



Configuration	h_2
Frozen	1.000
L4@10^{-4}	0.292
L6@10^{-4}	0.283
L8@10^{-4}	0.282
Full@5×10^{-5}	0.201
Full@10^{-4}	0.202

Figure 11: Image–text alignment h_2 (cosine to frozen text prototypes) on the slight-blur subset; bars and exact values side by side. Forgetting is already saturated at L4 and changes little at Full, so alignment forgetting cannot explain the selective collapse. Dashed line = frozen baseline.

structure *within the replayed fixed-boundary representation model* (direction is manipulated directly with norm held fixed); they do not establish causality for the original training dynamics, where head and encoder co-adapt. (vi) *Tolerance predicts the deployed-head outcome.* Although τ is computed entirely on the frozen probe, it predicts error under each configuration's own co-adapted deployed head. On Full@10^{-4} slight-blur (initially probe-correct, n=94), deployed error decreases monotonically across tolerance quintiles (78%, 75%, 22%, 0%, 0%; 0% at τ=∞; Table 16); every deployed error in this population occurs at $\tau \leq 1$ (precision 60%, recall 100%, AUC 0.96). The single-normal approximation $\tau_1 = m_0/[-q_{yk^*}]_+$ reaches only AUC 0.76, so competitor switching is predictively useful; pooled over all eight configurations (slight, n=752) the AUC of $-\tau$ is 0.89. This is not definitional—the deployed head is a different, jointly adapted boundary—and it transfers the frozen-boundary quantity to the actual failure (sharp has only one deployed error in this subset, too few to rank).

I Training-Distribution Controls

Table 17 reports two training-distribution controls at matched sample size (7,858, seed 42, η=10^{-4}, 5 epochs), evaluated on the common 11,232-image Bird test set; the *matched subsample* preserves the natural composition (338 sharp / 5,023 slight / 844 moderate / 1,653 unknown) and *no-slight* excludes all slight-blur images (leaving 2,117 sharp / 5,741 moderate). Three observations follow. First, excluding slight-blur images does not reproduce the selective signature: the model loses the unseen stratum outright (Full slight 37.7% vs. sharp 79.5%), which is a generic train–test mismatch rather than selective preservation. Second, the size-matched degraded subsample shows that fixed-epoch training on 1/6.7 of the data underfits regardless of composition (Full top-1 46.7%, L6 67.7%),

Table 12: Margin–drift quantities on the fixed linear-probe boundary (initially correct samples; endpoint margin excepted). flip_1: fraction flipped at α=1; α^*: median first-crossing dose among flipped; AUC: flip prediction by first-order score V; m_1^{ft}: median true-class margin at the deployed end point (own head), sh/sl.

	flip_1 (c0)		median α^* (flipped)		AUC of V		m_1^{ft} med
Configuration	sh	sl (ratio)	sh	sl	sh	sl	sh/sl (ratio)
Full@2×10^{-5}	0.117	0.191 (1.6)	0.72	0.47	0.49	0.59	6.77/5.58 (0.82)
Full@5×10^{-5}	0.068	0.266 (3.9)	0.72	0.34	0.38	0.70	7.43/5.21 (0.70)
Full@10^{-4}	0.087	0.266 (3.0)	0.57	0.35	0.45	0.74	7.62/5.05 (0.66)
L4@10^{-4}	0.175	0.255 (1.5)	0.86	0.66	0.55	0.62	5.83/5.50 (0.94)
L6@10^{-4}	0.136	0.223 (1.6)	0.84	0.64	0.48	0.68	6.66/6.00 (0.90)
L8@10^{-4}	0.175	0.287 (1.6)	0.87	0.70	0.57	0.63	6.65/5.80 (0.87)
L6@2×10^{-5}	0.204	0.298 (1.5)	0.81	0.78	0.51	0.62	5.26/4.86 (0.93)
L6@5×10^{-5}	0.165	0.277 (1.7)	0.78	0.67	0.52	0.68	6.22/5.76 (0.93)

Table 13: Flip fraction at α=1 under counterfactual drifts, Full@10^{-4} (random Q: mean over 20 rotations). Removing the single initial-boundary component ($\perp g$) does not remove crossings.

Counterfactual	sharp	slight	moderate
actual $z_0+\Delta z$	0.087	0.266	0.235
$\parallel g$ only	0.010	0.117	0.000
$\perp g$	0.243	0.383	0.392
$\perp g$, norm-matched	0.243	0.383	0.392
all norm-matched (r=15.28)	0.058	0.223	0.098
random $Q\Delta z$	0.777±0.079	0.701±0.073	0.878±0.106

removing the sample-size confound present in a naïve clean-vs-full comparison. Third, Full remains substantially below L6 on slight blur in *both* matched-size conditions (−15.6pp without slight-blur, −26.1pp with natural composition), so the high-LR Full disadvantage does not require degraded data and is not a data-size artifact; it is consistent with a generic high-LR full-adaptation instability on which the degradation-conditioned selective pattern is superimposed under full-data training. These controls are single-seed regime delimiters, not precision estimates.

J Dose Dynamics, Persistent Core, Toy Model, and Synthetic Control

This appendix gives the full numbers behind Sec. 6. The E=5/10/20 checkpoints are *independent runs* with separate cosine schedules (not checkpoints of one run); monotone patterns across all three families are emphasized because cross-budget increments mix schedule effects. Drift and fixed-boundary quantities use the same 600-image fixed subset (200 per stratum, seed 42); full-test accuracies use all 11,232 benchmark test rows (`data/fixed/test.csv`; the current `data/test.csv` additionally contains 197 synthetic-source rows tagged `unknown`, which belong to no evaluated stratum).

Persistent core and locking. On the slight-blur fixed subset the persistent (always-wrong at E=5/10/20) rate is 7–11% while sharp/moderate stay at 0%/0.5–1.0% (Table 20). Competitor locking is markedly depth-dependent: only 11.8–18.2% of Full's persistent errors share the same runner-up across all three budgets (47.1% at the lower LR share one wrong answer), versus 71.4%/92.9% for L6—Full wanders among competitors, L6 is pinned. Group characteristics (Table 21) show initial margins do not separate persistent from rescued samples (−0.92 vs. −1.39); the E=5→20 true-logit growth does (+0.18 vs. +5.12), while persistent samples' competitor logits grow (+2.49). Meanwhile own-head median margin grows with dose while own-head error plateaus (e.g. Full@5×10^{-5}: 5.19→10.25, acc 0.800→0.845; Table 19), so drifting confidence is not drifting correctness.

Table 14: Protection gap across all eight configurations (initially correct, fixed probe). P_{actual}, P_{random}: flip fractions at α=1; protection = $P_{\text{random}} - P_{\text{actual}}$ (paired rotation draws); paired gap uses the same 20 rotation draws.

	sharp			slight			paired
Configuration	P_{actual}	P_{random}	protect.	P_{actual}	P_{random}	protect.	gap $sh - sl$
Full@2×10^{-5}	0.117	0.753	0.637	0.191	0.666	0.475	0.162
Full@5×10^{-5}	0.068	0.770	0.702	0.266	0.705	0.439	0.264
Full@10^{-4}	0.087	0.777	0.689	0.266	0.701	0.435	0.254
L4@10^{-4}	0.175	0.634	0.460	0.255	0.544	0.288	0.171
L6@10^{-4}	0.136	0.634	0.499	0.223	0.542	0.319	0.180
L8@10^{-4}	0.175	0.640	0.466	0.287	0.545	0.258	0.208
L6@2×10^{-5}	0.204	0.632	0.428	0.298	0.537	0.239	0.189
L6@5×10^{-5}	0.165	0.635	0.470	0.277	0.525	0.248	0.221

Table 15: Margin-matched decision-projection tail, Full@10^{-4}: within each initial-margin bin, fraction with $q_{yk^*} < -m_0$ and realized flip fraction.

m_0 bin	stratum	n	$P(q < -m_0)$	flip
$(0, 0.5]$	sharp	28	0.000	0.179
	slight	21	0.143	0.381
$(0.5, 1]$	sharp	32	0.031	0.125
	slight	18	0.167	0.389
$(1, 2]$	sharp	21	0.000	0.000
	slight	19	0.211	0.263
$(2, \infty)$	sharp	22	0.000	0.000
	slight	36	0.028	0.139

M15: paired within-domain control on all 474 sharp benchmark photos. `scripts/eval_paired_synth_on_sharp.py` applies the exact Aircraft/Cars operators imported from `scripts/add_blur_degradations.py` (directional motion blur, defocus, camera shake, JPEG compression; slight/moderate; deterministic per-(image,type,level) seeds via crc32) to all 474 sharp rows of `data/fixed/test.csv`, producing 3,792 cached variants under `results/mechanism/paired_sharp_synth/`. The ten models are zero-shot CLIP and the nine dose checkpoints (Full@5×10^{-5}, Full@10^{-4}, L6@10^{-4}, E=5/10/20). Per-image outputs: `paired_sharp_synth_manifest.csv` (variants), `paired_sharp_synth_samples.csv` (37,920 prediction rows incl. `origin`=inat/unknown), `paired_sharp_synth_summary.csv` (accuracies and photo-cluster bootstrap CIs).

Depth and dose on synthetic motion. Paired slight-motion drops (pp) preserve the real-stratum depth ordering and partly close with dose: Full@10^{-4} 25.1/21.9/19.0, Full@5×10^{-5} 15.4/11.4/6.8, L6@10^{-4} 4.4/3.2/4.6 (E=5/10/20). Bootstrap 95% CIs at E=5: Full@10^{-4} motion 25.1 [21.3,28.9], pooled slight 6.3 [5.2,7.3]; L6@10^{-4} motion 4.4 [2.1,7.0], pooled slight 0.5 [−1.1,1.8]; the remaining types' CIs include 0 (full table in `paired_sharp_synth_summary.csv`). Synthetic moderate motion is catastrophic for every fine-tuned model rather than selective (accuracy: Full@10^{-4} 25.1/25.3/28.9%, Full@5×10^{-5} 34.4/31.8/40.6%, L6 40.5/42.6/39.0%, zero-shot 6.1%); the pooled-moderate loss in Table 22 is entirely attributable to motion—isotropic moderate cells differ from clean by $\leq$ 1.5pp at every checkpoint (e.g. Full@10^{-4}, E=5 defocus/shake/JPEG 97.7/97.9/97.9%).

Objective-strength calibration. `paired_sharp_laplacian.csv` reports Laplacian variance on 256-px bilinear center-resized grayscale images (lower = more blurred): sharp sources 1,992 (n=474); real test photos (500-photo seed-42 sample per stratum) slight 330 (IQR 127–704), moderate 13.8 (12.9–15.3); synthetic slight motion 392, camera shake 172, defocus 32, JPEG 2,055; synthetic moderate motion 85, shake 169, defocus 13.9, JPEG 2,106. Synthetic slight defocus is thus objectively as strong as the real moderate-stratum photos yet costs Full@10^{-4} 0.0pp; synthetic slight motion matches the real slight-stratum photos' strength and costs 25.1pp.

Table 16: Adaptation-tolerance calibration, Full@10^{-4} slight blur (initially probe-correct, n=94): quintiles of the exact multi-class τ (finite values) plus the τ=∞ group. Deployed error is under the configuration's own co-adapted head, not the frozen probe, so not definitional from τ.

τ bin	n	median τ	probe flip	deployed error
Q1 (lowest τ)	9	0.17	1.00	0.78
Q2	8	0.36	1.00	0.75
Q3	9	0.72	0.89	0.22
Q4	8	1.51	0.00	0.00
Q5 (highest τ)	9	6.55	0.00	0.00
τ=∞	51	∞	0.00	0.00

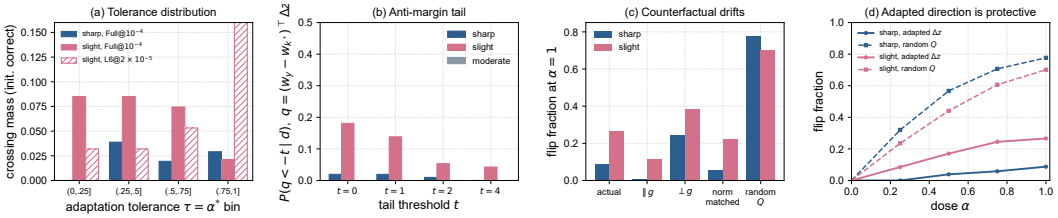


Figure 12: Counterfactual drift interventions on the fixed linear probe. **(a)** Crossing mass within adaptation-tolerance bins $\tau = \alpha^*$ (masses sum to flip fraction): Full@10^{-4} slight concentrates at small τ; safe L6@2×10^{-5} (hatched) keeps mass near τ=1. **(b)** Negative decision-projection tail $P(q_{yk^*} < -t \mid d)$, $t = 0,1,2,4$: slight dominates sharp and moderate at every t. **(c)** Flip fraction under the five counterfactual drifts (random Q: 20 draws). **(d)** Dose–response of actual vs. random drift (mean of 20 Q): the adapted direction is protective, far more for sharp than slight.

Photo-origin robustness. Of the 474 sharp rows, 159 are `inat` originals and 315 are rows tagged `unknown` whose paths point under `raw/synthetic/`. At Full@10^{-4}, E=5, paired slight-motion drops are 23.3pp on the `inat` subset (clean 95.6$\rightarrow$72.3%) and 26.0pp on the other subset (99.7$\rightarrow$ 73.7%), so the effect is not a property of either source tree.

Earlier lighter-library control (superseded). A separate, milder five-type chained library (17,360 rows / 3,472 bases, 202 held-out test bases; `data/raw/synthetic/metadata.csv`) showed near-zero synthetic losses for all fine-tuned checkpoints ($|\Delta| \leq 4.4$pp at both severities) with severely degraded zero-shot—a property of its under-calibrated operator strength and easy base selection, not a contradiction of the content-control claim. It is retained as a train/evaluation-distribution check only; per-image predictions and aggregates are in `results/mechanism/synthetic_control_{samples,summary}.csv`, and Table 22 is the basis for the Sec. 6 content-control conclusion.

Composition control (sample-size-matched, 3 conditions). To disentangle the natural slight-majority composition of the Bird train set (~67% slight) from fine-grained content as the driver of selectivity, we train Full and L6 at η=10^{-4}, E=5, seeds 42/1337/2024, on three matched conditions: (a) *natural(full)*—the full 52,968-row train set; (b) *natural(sub)*—a stratified 10,586-row subsample preserving the natural composition; (c) *synth-bal*—10,585 rows of balanced 1+2+2 composition (1 sharp + 2 slight + 2 moderate per base image) built from the 2,117 sharp bases via the same four operators as Table 22. Test evaluation uses the natural `test.csv` throughout. Two findings follow from Table 23. **(1) Balanced composition amplifies selectivity:** the synth-bal gap is 2$\times$ the natural-full gap, opposite to the hypothesis that balanced composition would remove selectivity; the gain-side decomposition shows synth-bal Full *negatively* transfers on slight ($G = -5.8$pp vs. zero-shot) while natural-composition Full retains a +19–+40pp gain, so the majority class (slight) is what the model is forced to learn, and removing that majority frees the optimizer to sacrifice slight for sharp/moderate. **(2) Sample size is not the driver:** halving the set 5$\rightarrow$1$\times10^4$ only mildly lowers S (+30 $\rightarrow$ +25), so the 2$\times$ jump under synth-bal is a composition effect. The Full$-$L6 depth penalty is stable across all three compositions (+9.3 to +10.8pp). Conclusion: the selectivity *base* is fine-grained content under high-LR Full; natural slight-majority composition is a buffer that modulates intensity, not the cause.

Table 17: Training-distribution controls (Bird, $\eta=10^{-4}$, seed 42, $E=5$): test accuracy (%) on the common 11,232-image test set. *Matched subsample*: equal-size random draw with natural composition; *No-slight*: all slight-blur images excluded.

Model	Train condition	N	Sharp	Slight	Moderate	Top-1
Full	full degraded	51,993	98.3	78.8	97.8	86.2
L6	full degraded	51,993	97.5	89.0	97.5	92.5
Full	matched subsample	7,858	53.2	43.6	50.4	46.7
L6	matched subsample	7,858	72.8	69.7	59.1	67.7
Full	no-slight	7,858	79.5	37.7	79.8	53.3
L6	no-slight	7,858	87.8	53.3	82.3	63.6

Table 18: Planned-budget dose sweep, full Bird test set (%) and fixed-subset drift. Full keeps drifting after accuracy plateaus; L6 drift saturates while accuracy keeps improving. PC1 alignment is $|\cos|$ of the centered drift PC1 at $E=10/20$ to $E=5$.

Config	E	Sharp	Slight	Top-1	$\|\Delta z\|$	$\lvert\cos_{\mathrm{PC1}}\rvert$
Full@5×10^{-5}	5	97.7	84.5	89.8	1.272	1.000
	10	99.0	85.5	90.6	1.286	0.822
	20	99.2	85.0	90.3	1.298	0.709
Full@10^{-4}	5	98.3	78.8	86.2	1.280	1.000
	10	99.0	80.4	87.3	1.296	0.860
	20	98.1	80.3	87.2	1.314	0.315
L6@10^{-4}	5	97.5	89.0	92.5	1.194	1.000
	10	98.7	89.4	93.0	1.194	0.946
	20	98.1	89.9	93.2	1.191	0.902

K Negative Results: DADU and TDRC

Figure 14 shows that the three strata (sharp, slight, moderate) produce nearly identical gate values at every λ: the router learns a layer-constant sparsity profile rather than a degradation-aware policy. Table 24 gives the exact diagnostic quantities.

Table 25 reports the TDRC (text-driven degradation-robust contrastive) objective applied to full-network adaptation. Relative to the matched Full baseline at the same $\eta=10^{-4}$ and seed, TDRC raises slight-blur accuracy (+6.4pp) at a comparable cost in sharp accuracy (−7.0pp): it narrows the selective gap by reallocating accuracy across strata, while top-1 (88.6%) remains below L6 partial adaptation (92.5%). The alignment-preserving loss therefore does not recover a safe full-adaptation configuration. The stratum-level shift was consistent across two independent repetitions of the TDRC run.

L Validation-Based Configuration Selection

Section 7 shows that neither tested training-time objective recovers a safe adaptation budget. As a simple post-training alternative, we evaluate whether *degradation-stratified validation* can screen out high-risk configurations using only the standard validation split, with no modification to training. Table 26 below gives the full validation and test columns. All 30 Bird sweep configurations (6 depths $\times$ 5 LRs, seed 42) are scored on the validation set (8,725 images with blur-level annotations: 446 sharp, 7,057 slight-blur, 1,222 moderate-blur). Each selection rule takes the argmax of a validation criterion over the 30 configurations; the selected configuration is then evaluated once on the held-out test set. Concretely, $\mathrm{R}_{\mathrm{clean}}$, $\mathrm{R}_{\mathrm{top1}}$, $\mathrm{R}_{\mathrm{deg}}$, and $\mathrm{R}_{\mathrm{min}}$ respectively select by validation sharp accuracy, validation top-1, validation slight-blur accuracy, and the minimum per-stratum validation accuracy. Selection uses validation per-stratum accuracy only: no test-set information is used in defining the rules or in choosing among configurations; each selected configuration is evaluated exactly once on held-out test, and all other test values appear only for reference.

Table 19: Fixed-boundary dose replay and own-head confidence on the slight-blur fixed subset (n_0=94 frozen-probe correct; n=200 own-head). $p(q<0)$: adverse decision-aligned projection; flip: crossings at α=1; rot: norm-matched random rotations; gap: rot−flip protection.

Config	E	$p(q<0)$	flip	rot	gap	own-head acc / median margin
Full@5×10^{-5}	5	0.117	0.266	0.705	0.439	0.800 / 5.19
	10	0.106	0.202	0.736	0.534	0.865 / 8.16
	20	0.106	0.213	0.760	0.547	0.845 / 10.25
Full@10^{-4}	5	0.181	0.266	0.701	0.435	0.795 / 5.05
	10	0.128	0.255	0.751	0.495	0.800 / 6.90
	20	0.138	0.245	0.793	0.548	0.820 / 9.54
L6@10^{-4}	5	0.202	0.223	0.542	0.319	0.850 / 5.99
	10	0.138	0.213	0.546	0.333	0.880 / 7.28
	20	0.181	0.223	0.540	0.316	0.885 / 8.59

Table 20: Persistent-core group counts (slight-blur fixed subset, n=200/config; groups from own-head correctness at E=5/10/20: persistent/rescued/stable/late-emerging/transient) with competitor locking among persistent errors (RU: same runner-up; WA: identical wrong answer, all three budgets). CIs: 2,000-resample bootstrap.

Config	pers.	resc.	stable	late	trans.	pers. rate [CI]	same RU (n)	same WA
Full@5×10^{-5}	17	14	150	10	9	0.085 [0.050,0.125]	0.118 (17)	0.471
Full@10^{-4}	22	9	143	16	10	0.110 [0.065,0.160]	0.182 (22)	0.227
L6@10^{-4}	14	8	160	10	8	0.070 [0.035,0.105]	0.714 (14)	0.929

The three rules that use aggregate or degradation-aware validation criteria all select the same configuration, L8@2×10^{-5}, which has the best test top-1 among the 30 sweep configurations while avoiding the high-risk regime (−7.6pp vs. −19.5pp for Full@10^{-4}). The one rule that fails is selection by clean-stratum validation accuracy alone: it retains a full-network configuration (Full@5×10^{-5}) whose test selective gap (−13.2pp) remains large. Degradation-stratified validation can thus serve as a simple model-selection criterion for avoiding high-risk configurations. We emphasize the boundary of this claim: it does not resolve the safe adaptation budget in general, and its reliability inherits the degradation-stratum coverage of the validation split. A two-domain replication, including the cases where even clean-only selection succeeds, is reported next.

Cross-domain replication on Aircraft and Cars. These benchmarks ship a single degraded evaluation file (FGVC-Aircraft: 3,333 images; Stanford Cars: 8,041 images; each image at sharp/slight/moderate blur) with no separate validation images. We therefore apply a *pre-registered image-level 50/50 split*: images are ordered by image id and assigned to the selection/held-out halves by sorted-position parity, with all three blur rows of an image kept in the same half. This yields 5,001/4,998 rows (Aircraft) and 12,063/12,060 rows (Cars), each stratum balanced 50/50 by construction. The candidate set is a 12-configuration subgrid (L4/L6/L8/Full × $\{2\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$, seed 42): all 12 are scored on the selection half, the four rules select from that half alone, and selected/reference configurations are each evaluated once on the held-out half, which never enters any selection decision.

Two findings differ from Bird. First, *all four rules agree*, including clean-only R_{clean}: unlike Bird, where Full@5×10^{-5} preserves sharp accuracy (99.1% validation) and is exposed only by stratified criteria, high-dose Full on Aircraft/Cars also degrades sharp accuracy, so the high-risk corner is visible even without stratification. Stratified criteria are indispensable specifically when the failure is stratum-selective, as on Bird. Second, the avoided loss is primarily aggregate top-1 (+9.4/+11.7pp), whereas the selective gaps differ only modestly (within 0.3–0.5pp). The replication thus confirms that the protocol avoids the pre-specified high-risk corner across domains, but the strongly stratum-selective signature itself remains Bird-specific in magnitude. Boundaries: seed 42 only, a 12-configuration subgrid rather than the full 30-configuration sweep, and the same blur pipeline in both halves (no cross-corruption-type transfer).

Table 21: Pooled group characteristics over the three dose families (slight blur): persistent n=53, rescued n=31, stable n=453 (medians; bootstrap CIs in source CSV). Locking rows: same runner-up at all three budgets (all groups); identical wrong answer (persistent only).

Quantity	persistent	rescued	stable
frozen-probe m_0	−0.92	−1.39	+0.10
zero-shot m_0	−0.015	−0.016	−0.010
$\Delta\ell_{\text{true}}$ (E=5→20)	+0.18	+5.12	+4.17
$\Delta\ell_{\text{comp}}$	+2.49	−0.20	+0.17
margin at E=5	−2.35	−1.16	+6.48
margin at E=20	−4.53	+3.73	+9.81
same runner-up all 3	0.30	0.23	0.17
same wrong answer all 3	0.49	—	—

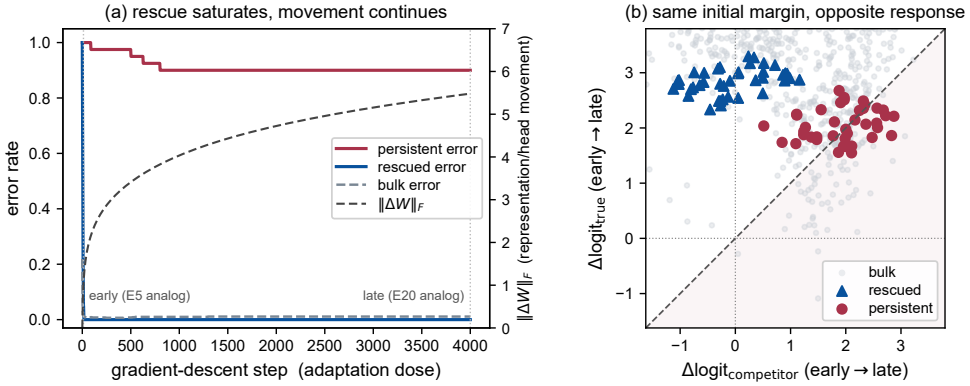


Figure 13: **Minimal existence model** (2-D three-class softmax logistic regression, full-batch GD, seed 42). Bulk: 900 isotropic Gaussian samples (300/class); the bulk-supported head W^* is fit first and W_0 is its boundary rotated by 80°. Two degraded groups of 40 class-0 points are nearest-neighbor-matched on signed initial margin under W_0 (median −0.401 both): rescued points end on the correct side of W^*; persistent points sit in the competitor-class bulk. **(a)** Rescue error hits 0 by step 12; persistent error stays at 0.90 through step 4000 while $\|\Delta W\|_F$ grows to 5.5. **(b)** Logit growth E=5→20 analog: rescued true-class logits grow while competitor logits shrink; for persistent points both grow together (+2.02 vs. +1.98).

M Reproducibility Details

Hardware and software. Experiments were conducted on NVIDIA A100 GPUs (40GB memory) using PyTorch 2.6.0 with CUDA 12.4. The codebase uses OpenCLIP 2.24 for model loading and standard torchvision datasets. No specialized hardware or proprietary libraries were used beyond those listed above.

Random seeds. Three seeds were used: 42, 1337, 2024. All seeds were set via both `torch.manual_seed(seed)` and `random.seed(seed)` at the top of each training script. Weight initialization follows the default OpenCLIP initialization; seeds affect data order augmentation randomness and batch sampling order.

Training command template. The standard command for a partial-adaptation run is:

```
python run_formal.py --dataset bird --depth L6 --lr 1e-4 --seed 42 --workers 0
```

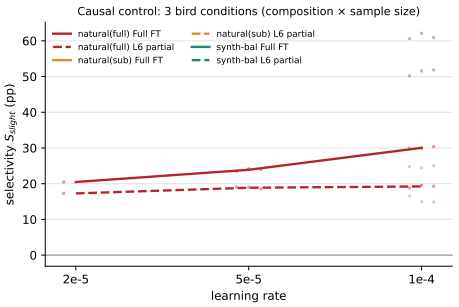
For cross-domain runs:

```
python run_cross_domain.py --dataset aircraft --depth L8 --lr 5e-5 --seed 1337
```

All hyperparameters (epochs=5, batch=32, AdamW optimizer, cosine schedule) are defaults in the scripts unless overridden.

Table 22: **M15 paired control.** Accuracies (%) on clean sharp photos and synthetic variants of the same photos (474 photos × 4 types; slight per type and pooled, plus pooled moderate). Paired clean−variant Δ with photo-cluster bootstrap 95% CIs: see text.

	slight synthetic variants						moderate
Checkpoint	clean	motion	defocus	shake	JPEG	pooled	pooled
zero-shot	28.3	15.4	16.2	19.0	23.0	18.4	10.4
Full@5×10^{-5}, E=5	97.7	82.3	98.3	98.1	98.3	94.3	82.0
Full@5×10^{-5}, E=10	98.9	87.6	98.5	98.9	98.9	96.0	81.7
Full@5×10^{-5}, E=20	99.2	92.4	98.5	98.7	99.4	97.3	83.8
Full@10^{-4}, E=5	98.3	**73.2**	98.3	98.1	98.5	92.0	79.6
Full@10^{-4}, E=10	98.9	77.0	98.1	97.9	98.5	92.9	79.8
Full@10^{-4}, E=20	98.1	79.1	98.1	98.3	98.3	93.5	80.7
L6@10^{-4}, E=5	97.5	93.0	98.7	97.7	98.5	97.0	82.9
L6@10^{-4}, E=10	98.7	95.6	98.5	98.5	98.5	97.8	83.8
L6@10^{-4}, E=20	98.1	93.5	98.3	98.3	99.4	97.4	83.1



	n_{train}	nat-full 52,968	nat-sub 10,586	s-bal 10,585
S_{slight}	Full	$+30.0\pm0.3$	$+24.7\pm0.3$	$+61.2\pm0.8$
	L6	$+19.2\pm0.4$	$+15.5\pm0.9$	$+51.2\pm0.9$
G_{sharp}	Full	+70.5	+43.9	+55.4
G_{slight}	Full	+40.4	+19.1	−5.8
G_{moderate}	Full	+87.3	+61.3	+52.9

Table 23: **Composition control** (3 conditions, E=5, $\eta=10^{-4}$, 3 seeds). *Left*: selectivity S_{slight} under natural-full (red), natural-sub (orange), and synth-balanced (green) composition; Full FT (solid) vs. L6 partial (dashed). *Right*: S (zero-shot-anchored, pp) and G (gain over zero-shot, pp).

Anonymous code release. The full codebase including training scripts, evaluation pipelines, and result-processing utilities will be released in an anonymized repository upon acceptance; the complete code is provided in the supplementary materials. No author identities, institutional affiliations, or personal paths appear in any file.

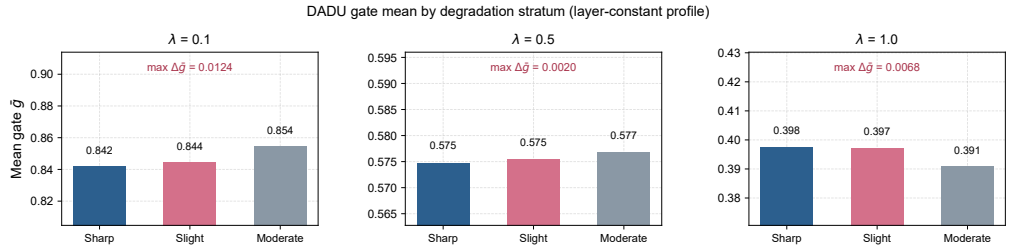


Figure 14: DADU mean gate $\bar{g}$ by degradation stratum at three budget weights λ. The sharp/slight/moderate bars are nearly identical at every λ (max $\Delta\bar{g} < 0.01$), confirming the router collapses to a layer-constant sparsity profile rather than degradation-conditioned routing.

Table 24: DADU gate diagnostic sweep. Higher λ reduces mean gate (more sparse) but does not produce degradation-conditioned routing.

λ	d_{eff} (sh/sl/mod)	Δ_g	std(g)	ρ
0.1	5.05/5.07/5.13	+0.012	0.004	+0.540
0.5	3.45/3.45/3.46	+0.002	0.001	+0.444
1.0	2.39/2.38/2.35	−0.007	0.002	−0.499

Table 25: TDRC objective (Full network, Bird, $\eta = 10^{-4}$, seed 42): stratum accuracies and pp changes vs. baseline Full fine-tuning.

Objective	Sharp	Slight-blur	Δ Sharp	Δ Slight
Baseline Full FT	98.3	78.8	—	—
Full FT + TDRC	91.4	85.2	−7.0	+6.4

Table 26: Full validation-based selection table (Bird, seed 42). Gap = sharp−slight (pp); validation accuracies as fractions.

Rule	Selected config.	Val.			Test			
		sharp	slight	top-1	sharp	slight	top-1	gap
R_{clean} (sharp)	Full@5×10^{-5}	0.991	0.841	0.869	97.7	84.5	89.8	−13.2
R_{top1} (top-1)	L8@2×10^{-5}	0.971	0.888	0.904	97.3	89.7	93.0	−7.6
R_{deg} (slight)	L8@2×10^{-5}	0.971	0.888	0.904	97.3	89.7	93.0	−7.6
R_{min} (worst-stratum)	L8@2×10^{-5}	0.971	0.888	0.904	97.3	89.7	93.0	−7.6
(Avoided high-risk)	Full@10^{-4}	0.991	0.774	0.815	98.3	78.8	86.2	−19.5