WHAT FIT HIDES: THE GEOMETRY OF FEW-SHOT SUBPOPULATION REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Two model updates can produce nearly identical changes in fixed edit-loss measurements yet generalize differently to the wider target subpopulation. We study this ambiguity in few-shot subpopulation repair, where a model is edited using a small correction set. A local geometric analysis quantifies how much held-out target improvement can remain ambiguous given identical or approximately matching edit measurements. With additional target, non-target, and reference probes, we construct a diagnostic update that achieves a specified local target improvement while minimizing a regularized non-target representation drift cost within the edit-gradient span. The same geometry predicts its locality advantage over the projected target-gradient update. Across six CIFAR-100 model runs, edit-loss differences shrink approximately quadratically with update size, while held-out target-score differences shrink approximately linearly. Using the same probes and matching validation improvement, the diagnostic achieves 16.1–18.5% higher model-median test target gain per unit off-target drift than the projected target gradient. Validation geometry predicts the projected-gradient-to-diagnostic test drift ratio without fitted calibration, with 4.31% median error relative to the excess ratio above one, compared with 22.04% for a leave-one-model-out constant predictor. The locality advantage remains with fewer probes and an additional edit-loss descent constraint. Paired continuations show that updates with the same immediate target improvement can have different effects after training resumes. Waterbirds and controlled ImageNet-20 reproduce the generalization and locality distinctions under a natural subgroup failure and with larger ResNet and ViT models.

1 INTRODUCTION

Suppose a deployed model performs poorly on a small but meaningful subpopulation. A natural response is to collect a few examples, edit the model on those examples, and check that the edit loss went down. This test answers one question: did the model fit the examples it saw? It does not establish whether the rest of the target subpopulation improved, whether the change will survive continued training, or whether unrelated behavior changed.

The three broader requirements—generalization to held-out target examples, locality on non-target examples, and retention under resumed training—are easy to confuse with one another during training. Few-shot post-hoc edits show how they differ. The edit set is small, the model is overparameterized, and each edit changes parameters shared by many subpopulations. Two updates can therefore look the same on the edit set yet produce substantially different held-out outcomes. This is a local form of underspecification (D'Amour et al., 2022). One update may also cause more off-target drift than another, and updates with the same immediate benefit may evolve differently after training resumes. This concern is especially relevant for continuously updated systems, in which an edit must remain useful after the model returns to its ordinary training stream.

We formalize this setting with three sets. The edit examples $\mathcal{E}$ provide the correction signal and may include target examples, clean anchors, and comparison examples. The set $\mathcal{U}$ contains target examples outside the edit set. We construct the diagnostic using validation target probes and evaluate

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.

the frozen update on separate test examples. Target behavior should improve, while behavior on the non-target set $\mathcal{P}$ should remain stable. We call this preservation *locality*.

In model-editing terminology, edit success is reliability, held-out target improvement is generalization, and non-target preservation is locality (Mitchell et al., 2022a;b; Yang et al., 2024). Retention measures persistence under resumed training. Our analysis separates what edit-set fit can establish about generalization, what local geometry permits for locality, and what immediate gains imply about retention.

Our analysis gives three quantities for assessing repair. Residual target sensitivity measures what fixed edit losses leave undetermined, including when measurements match only approximately. The minimum-drift cost gives the achievable local tradeoff for a required target improvement within the edit-gradient span. Its comparison with the projected target gradient predicts the value of using non-target geometry under the same information budget. We test these quantities on the nonlinear models and evaluate validation-based predictions on separate test examples. Paired continuations also test whether matching immediate improvement determines later effects. Here *few-shot* refers to the correction set; constructing the diagnostic requires additional probes.

We use Waterbirds and controlled ImageNet-20 to test whether these generalization and locality distinctions extend beyond the CIFAR construction. These settings include a natural subgroup failure and larger convolutional and transformer models.

2 Problem Setup: Few-Shot Subpopulation Repair

We distinguish what an editing procedure can change, what its edit examples measure, and how the resulting update affects held-out target and non-target examples.

Let θ denote the model parameters, θ_0 a degraded checkpoint, and $\mathcal{V}$ the linear space of allowed parameter updates; all gradients and Jacobians below are restricted to $\mathcal{V}$. We call a correction procedure that selects an update in $\mathcal{V}$ an *editor*. In our primary setting, this procedure may change the trainable ("online") encoder and its projection head. Let $r_{\mathcal{U}}(\theta)$ be a smooth held-out target score, with larger values indicating better performance, and let $z_{\mathcal{P}}(\theta)$ collect the representations of the non-target examples. The edit set $\mathcal{E}$ supplies correction supervision, while $\mathcal{U}$ contains target examples outside the edit set. The set $\mathcal{P}$ contains non-target examples whose representations we seek to preserve.

We collect the measurements made on the edit examples into a differentiable vector $C_{\mathcal{E}}(\theta) \in \mathbb{R}^m$, where m is the number of fixed measurements. Its entries need not be model predictions. In our CIFAR-100 experiment, they are the same $m = 12$ stochastic edit losses whose gradients define the initial locality-comparison span. At θ_0, define

$$A_{\mathcal{E}} \coloneqq DC_{\mathcal{E}}(\theta_0), \qquad a_{\mathcal{U}} \coloneqq \nabla r_{\mathcal{U}}(\theta_0). \tag{1}$$

Here D denotes the Jacobian. First-order effects are the linear predictions for a sufficiently small parameter update. The rows of $A_{\mathcal{E}}$ are the edit-loss gradients. Their span is the subspace observable from the edit measurements: it contains the parameter directions that linear combinations of those measurements can detect to first order.

For the locality comparison, we restrict updates to the fixed edit-gradient span $\mathcal{S}_{\mathcal{E}} = \operatorname{range}(A_{\mathcal{E}}^{\mathsf{T}}) \subseteq \mathcal{V}$. The edit measurements are injective on this span: $\mathcal{S}_{\mathcal{E}} \cap \ker A_{\mathcal{E}} = \{0\}$. Thus exact first-order ambiguity concerns the larger space $\mathcal{V}$, not two distinct span-restricted updates. The nonzero witness of this ambiguity lies outside $\mathcal{S}_{\mathcal{E}}$. Multistep learned updates can also extend outside this span. The retention alternative includes a component outside the span as well. Only the locality comparators are all restricted to $\mathcal{S}_{\mathcal{E}}$.

For locality, let $J_{\mathcal{P}} = Dz_{\mathcal{P}}(\theta_0)$ be the non-target representation Jacobian and let $W_{\mathcal{P}} \succeq 0$ be a fixed weighting matrix. The local off-target drift cost of $v \in \mathcal{V}$ is

$$d_{\mathcal{P}}(v)^2 \coloneqq v^{\mathsf{T}} J_{\mathcal{P}}^{\mathsf{T}} W_{\mathcal{P}} J_{\mathcal{P}} v. \tag{2}$$

With averaging weights, $d_{\mathcal{P}}$ is the root-mean-square (RMS) off-target drift; squaring changes before averaging prevents cancellation. Non-target accuracy, neighbor purity, and effective rank complement this geometric measure with behavioral and representation readouts (Appendix B.2).

3 LOCAL GEOMETRY OF SUBPOPULATION REPAIR

Our analysis quantifies remaining target ambiguity, the minimum local drift at a required target gain, and the benefit of using the drift metric. These quantities give tests on the original nonlinear model. A final identifiability result addresses effects after continued training. Proofs are in Appendix A.

3.1 EDIT-SET IDENTIFIABILITY

Edit-set fit determines first-order held-out target improvement if and only if every update direction invisible to the edit losses is also irrelevant to that improvement. To express this condition, let $P_{\mathcal{E}}$ be the orthogonal projector onto $\mathrm{range}(A_{\mathcal{E}}^{\mathsf{T}})$. This projector keeps the part of a parameter direction that the edit measurements can detect and removes the rest.

Theorem 1 (Local edit-set identifiability)**.** *The following statements are equivalent:*

$$\begin{aligned} &\text{(i)} \quad A_{\mathcal{E}} v_1 = A_{\mathcal{E}} v_2 \Longrightarrow a_{\mathcal{U}}^{\mathsf{T}} v_1 = a_{\mathcal{U}}^{\mathsf{T}} v_2 \quad \textit{for every } v_1, v_2 \in \mathcal{V}; \\ &\text{(ii)} \quad \ker(A_{\mathcal{E}}) \subseteq \ker(a_{\mathcal{U}}^{\mathsf{T}}); \\ &\text{(iii)} \quad a_{\mathcal{U}} \in \mathrm{range}(A_{\mathcal{E}}^{\mathsf{T}}); \\ &\text{(iv)} \quad (I - P_{\mathcal{E}}) a_{\mathcal{U}} = 0. \end{aligned} \tag{3}$$

When this condition fails, the missing target-relevant component gives a concrete counterexample. If $q = (I - P_{\mathcal{E}}) a_{\mathcal{U}} \neq 0$ and $\widehat{q} = q / \|q\|$, then $A_{\mathcal{E}} \widehat{q} = 0$ while $a_{\mathcal{U}}^{\mathsf{T}} \widehat{q} = \|q\|$. For any base direction d, the pair $d + \widehat{q}$ and $d - \widehat{q}$ is first-order equivalent under the defining edit measurements but not on $\mathcal{U}$.

If $C_{\mathcal{E}}$ and $r_{\mathcal{U}}$ have locally bounded second derivatives, perturbing the two directions by an update-size multiplier ϵ gives

$$\|C_{\mathcal{E}}(\theta_0 + \epsilon(d + \widehat{q})) - C_{\mathcal{E}}(\theta_0 + \epsilon(d - \widehat{q}))\| = O(\epsilon^2), \tag{4}$$

$$r_{\mathcal{U}}(\theta_0 + \epsilon(d + \widehat{q})) - r_{\mathcal{U}}(\theta_0 + \epsilon(d - \widehat{q})) = 2\epsilon \|q\| + O(\epsilon^2). \tag{5}$$

Under these smoothness assumptions the edit gap is $O(\epsilon^2)$, whereas the target gap has a nonzero linear term. An exact exponent of two also requires a nonzero quadratic coefficient. Exponents measured at finite update sizes show how closely the observed scaling follows this prediction. The guarantee concerns the fixed measurements in $C_{\mathcal{E}}$; independent augmentation draws provide a further test of generalization beyond those measurements. Figure 1A tests the predicted contraction.

Appendix B.11 allows the edit measurements to differ within a finite tolerance. It gives the largest possible target disagreement for specified bounds on measurement differences and update radius. It also gives an upper bound for finite updates when local second derivatives are bounded.

3.2 MINIMUM-DRIFT REPAIR

We next ask whether a specified first-order held-out target improvement can be reached inside the edit-gradient span while minimizing off-target representation drift and update size. Let the columns of Q form an orthonormal basis for $\mathcal{S}_{\mathcal{E}}$, write $v = Qw$, and define

$$b \coloneqq Q^{\mathsf{T}} a_{\mathcal{U}}, \qquad H \coloneqq Q^{\mathsf{T}} J_{\mathcal{P}}^{\mathsf{T}} W_{\mathcal{P}} J_{\mathcal{P}} Q, \qquad H_\lambda \coloneqq H + \lambda I, \tag{6}$$

where $\lambda > 0$. The ridge term λI makes the solution unique and discourages large coefficients in weakly measured directions. For a required first-order held-out target improvement $\rho > 0$, consider the convex quadratic program (Boyd & Vandenberghe, 2004)

$$\min_{w} \; w^{\mathsf{T}} H_\lambda w \quad \text{subject to} \quad b^{\mathsf{T}} w \geq \rho. \tag{7}$$

Theorem 2 (Regularized minimum-drift update)**.** *Positive held-out target improvement is feasible in $\mathcal{S}_{\mathcal{E}}$ if and only if $b \neq 0$. When it is feasible, the unique optimizer and minimum cost are*

$$w^\star = \frac{\rho H_\lambda^{-1} b}{b^{\mathsf{T}} H_\lambda^{-1} b}, \qquad D_{\mathcal{E},\lambda}^\star(\rho) = \frac{\rho^2}{b^{\mathsf{T}} H_\lambda^{-1} b}. \tag{8}$$

The minimum-cost direction $QH_\lambda^{-1}b$ has the metric-preconditioned form of natural-gradient and Gauss–Newton updates (Martens, 2020), with representation drift defining the metric. We use it as a *diagnostic update*. The objective constrains target gain and penalizes representation drift and update size. Appendix B.14 adds an edit-loss descent constraint and tests the remaining locality advantage. Finite-step comparisons match measured validation target gain within a radius where target and non-target linear approximations remain accurate. For a reachable coefficient vector w, we summarize the local tradeoff by

$$\eta_\lambda(w) \coloneqq \frac{b^\mathsf{T} w}{\sqrt{w^\mathsf{T} H_\lambda w}}, \qquad \eta_0(w) \coloneqq \frac{b^\mathsf{T} w}{\sqrt{w^\mathsf{T} H w}}. \tag{9}$$

These are the regularized and unregularized first-order *repair-to-drift ratios*; η_0 requires positive drift. Finite-update gain/RMS instead uses measured target gain and unregularized RMS. Larger positive values are better.

The value of the metric can itself be quantified. At gain ρ, the projected target gradient has coordinates $w_b = \rho b/(b^\mathsf{T} b)$, and

$$\frac{w_b^\mathsf{T} H_\lambda w_b}{D^\star_{\mathcal{E},\lambda}(\rho)} = \frac{(b^\mathsf{T} H_\lambda b)(b^\mathsf{T} H_\lambda^{-1} b)}{(b^\mathsf{T} b)^2} \geq 1. \tag{10}$$

Equality holds exactly when b lies in one eigenspace of H_λ. Thus the advantage depends on how target sensitivity spans directions with different drift costs. This squared regularized ratio is distinct from nonlinear test RMS; Appendix B.10 gives the corresponding unregularized prediction, tested in Figure 1B.

The unregularized analysis identifies a special case with first-order target improvement and only second-order off-target drift. Suppose a direction in the edit-gradient span improves the target score to first order and lies in the null space of the non-target representation Jacobian. Then Corollary 1 gives $\Theta(\epsilon)$ held-out target improvement but only $O(\epsilon^2)$ off-target RMS drift. Thus off-target drift vanishes at least quadratically while useful repair remains first order. This uses the same null-space reasoning as Theorem 1, with the non-target representation Jacobian replacing the edit-measurement Jacobian. Section 5.2 evaluates the finite-update repair–locality comparison.

The construction uses a held-out target-subpopulation gradient to form b and a non-target set to form H. These additional probes let the diagnostic assess what the local geometry permits. A procedure using only the edit set has less information. If the diagnostic update beats a learned update, then the local geometry allows less off-target drift than the learned update produces. Experiments with finite probe sets in Section 5 test how much data are needed to estimate it.

3.3 Retention under Resumed Training

The final question is whether two updates with equal immediate held-out target-score effects must remain equivalent after ordinary training resumes. Continued training evolves the complete state s: model parameters and auxiliary model/optimizer state.

Let $I(v)$ denote the training state obtained by applying parameter update v while leaving the other state components unchanged. Fix randomness ω to specify the training examples, order, augmentations, and other stochastic choices shared by the edited and unedited runs. Write R_T for the resulting T-epoch training map, omitting the fixed ω from the notation. When applied to a training state, $r_\mathcal{U}$ uses that state's model parameters. Define the intervention's counterfactual effect after T epochs by

$$\Delta_T(v) \coloneqq r_\mathcal{U}(R_T(I(v))) - r_\mathcal{U}(R_T(I(0))). \tag{11}$$

The second term follows training with the same data and randomness, starting without an intervention. This paired no-intervention baseline accounts for changes caused by ordinary training alone. A comparison with the static degraded checkpoint would mix these changes with the effects of retaining the edit. Define $a_0 \coloneqq \nabla_v \Delta_0(v)|_{v=0} = a_\mathcal{U}$ and $a_T \coloneqq \nabla_v \Delta_T(v)|_{v=0}$ as the immediate and future target sensitivities to the intervention. On a smooth branch with locally bounded second derivatives,

$$\Delta_0(v) = a_0^\mathsf{T} v + O(\|v\|^2), \qquad \Delta_T(v) = a_T^\mathsf{T} v + O(\|v\|^2). \tag{12}$$

Theorem 3 (Local identifiability of future effects)**.** *Equal immediate first-order held-out target-score effects determine equal future first-order held-out target-score effects for every pair $v_1, v_2 \in \mathcal{V}$ if and only if*

$$\ker(a_0^\mathsf{T}) \subseteq \ker(a_T^\mathsf{T}), \quad \textit{equivalently} \quad a_T = c_T a_0 \tag{13}$$

for some scalar c_T.

Thus equal immediate effects determine equal future effects only when the future and immediate sensitivities are proportional. For two fixed directions $d_A, d_B \in \mathcal{V}$ with $a_0^\top d_B \neq 0$, choose $c = (a_0^\top d_A)/(a_0^\top d_B)$ and set $h = d_A - cd_B$. Then $a_0^\top h = 0$. If $a_T^\top h \neq 0$, the matched pair has an immediate gap of $O(\epsilon^2)$ and a future gap of $O(\epsilon)$. Theorem 1 applies the identifiability criterion to edit measurements and held-out effects; here it applies to immediate and future effects. Section 5.3 tests this prediction and reports the finite-update comparison in Table 1.

Discrete top-k selections can change abruptly, so the continued-training dynamics are only piecewise differentiable. We test the theorem's smooth prediction at progressively smaller update sizes and report finite-size comparisons separately.

4 Experimental Design

CIFAR experiments. Each theory test starts from a degraded CIFAR-100 checkpoint (Krizhevsky, 2009). The edit set $\mathcal{E}$ contains 60 examples: 30 positive examples drawn from clean target-class exemplars and target-subpopulation examples, and 30 comparison negatives. The set $\mathcal{U}$ contains target examples outside the edit set, while $\mathcal{P}$ contains non-target examples from 90 unrelated classes. We measure target improvement using the mean similarity margin on $\mathcal{U}$: the difference in mean cosine similarity to the target and comparison reference groups. Normalized recovery reports the percentage of the task-specific gap to the reference level recovered; values above 100% indicate that this level was exceeded.

The theory tests use six stochastic training runs that share a pretrained initialization and data split. Each run is evaluated under three nested allocations that vary the balance of clean exemplars and target-subpopulation examples (Appendix B.2), giving eighteen model–allocation *configurations*. The six model runs are the replication units; comparisons and radius sweeps are nested within them (Appendix B.15). In each configuration, 12 fixed stochastic edit-loss gradients define the locality-comparison span, and the locality directions differ only in how they combine these gradients. We express update sizes relative to the norm of the corresponding 20-epoch reference update and test all three theoretical distinctions on the original nonlinear model. Locality is also evaluated on test images outside diagnostic construction (Appendix B.5). We report effect ranges across model runs and configurations.

The broader CIFAR study varies edit composition, strength, task, and reference group. The reference group is the semantic group toward which the correction should move the target examples. We use supervised contrastive learning (Khosla et al., 2020) and prototype cross-entropy objectives. "Aquatic" and "Flower" denote target-subpopulation tasks within the same dataset and architecture.

Additional datasets and architectures. Waterbirds uses waterbird-on-land as a naturally failing target subpopulation $\mathcal{U}$ and the remaining groups as non-target (Sagawa et al., 2020). ImageNet-20 combines the Imagenette and Imagewoof subsets (Howard, 2019) of ImageNet (Deng et al., 2009), using pretrained ResNet-50 and ViT-B/16 (He et al., 2016; Dosovitskiy et al., 2021). Controlled mixing of target-class classifier rows isolates target-output corruption while preserving inputs and the pretrained backbone. The natural-adaptation exclusion and corruption strengths are reported in Appendix B.16. Complete protocols, comparison rules, and radius windows are given in Appendix B; sensitivity checks are described in Appendix B.5.

5 Results

We first report the three theorem-matched tests, then assess generalization and locality across additional datasets and architectures.

5.1 Edit Fit Does Not Determine Target Generalization

Standard learned edits already separate edit fit from target generalization. In the aquatic CIFAR task, target-subpopulation-only edit sets can reach perfect edit-set accuracy while normalized recovery

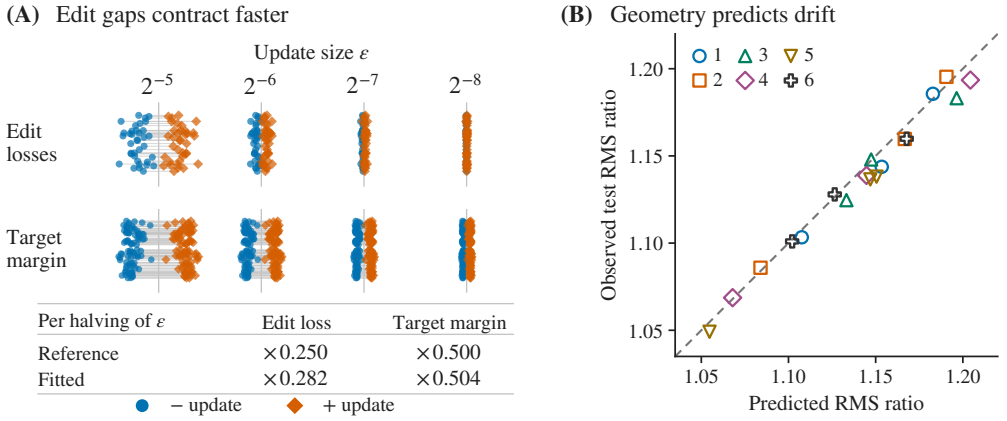


Figure 1: **A:** Paired edit-loss and held-out target-margin responses across six models, midpoint-centered with a separate fixed scale per row. The same responses appear at each update size; the table compares smooth-reference and fitted contraction factors (the quadratic reference assumes a nonzero leading term). **B:** Validation-geometry predictions and observed test RMS drift ratios (projected target / diagnostic); the dashed line denotes equality. Each point summarizes four validation-matched gain levels. Shapes identify six models, with three allocations nested within each. Predictions use no fitted calibration. Reported relative errors use the excess ratio above one (Appendix B.10).

remains near zero or negative. Changing the reference group can reverse the ordering suggested by edit fit, and the useful balance of clean exemplars and target examples varies across tasks (Appendix B.2).

Completed learned updates extend beyond the initial twelve-gradient span: their orthogonal components contain 91.2–93.6% of update norm and have nonzero target-score derivatives of either sign. Small-scale nonlinear probes confirm target effects outside the span (Appendix B.8). This connects the allowed-space analysis to learned editing, without implying that ordinary editors produce exactly matching edit measurements. To test identifiability separately, we hold the 12 edit measurements fixed and test the paired-direction construction of Theorem 1 (Figure 1A).

The 12 fixed CIFAR edit-loss gradients capture a mean 30.1% of the held-out target-gradient norm, measured as $\|P_{\mathcal{E}} a_{\mathcal{U}}\|/\|a_{\mathcal{U}}\|$. The mean squared-norm fraction is 9.26%. The remaining component is invisible to the edit measurements to first order. We test the pair constructed from this missing component over a radius window from $1/256$ through $1/32$ of the reference-update norm. If a separation scales as ϵ^p, then $p = 2$ is quadratic and $p = 1$ is linear. The root-mean-square difference between the two vectors of 12 defining edit losses has mean log–log exponent 1.83, compared with 0.99 for the held-out target-score gap. All eighteen edit-gap exponents are closer to quadratic than linear: edit-loss gaps contract faster than target-score gaps, as predicted by Equations (4)–(5).

This separation is specific to the fixed edit measurements used to construct the pair. Their first-order differences cancel to numerical precision. An independently drawn augmentation outside these measurements does not show the same quadratic contraction. Increasing the number of defining draws from 12 to 48 on the same edit images raises mean target-gradient norm coverage from 30.1% to 50.9%. The remaining unmeasured component still produces a positive target separation on the independent test set in all eighteen configurations (Appendix B.12). Full scaling statistics and display details are given in Appendices B.5 and B.3.

5.2 Equal Improvement with Less Off-target Drift

We compare off-target drift inside the same edit-gradient span at equal measured improvement on held-out target examples. A lower-drift update exists in all eighteen CIFAR configurations.

In all eighteen configurations, the diagnostic direction improves the held-out target score to first order, whereas descent along the negative mean edit-loss gradient decreases it. Their mean unregularized first-order ratios are 0.087 and -0.019, respectively. A small step along the completed fine-tuning update also lowers the target score, although the full update improves it.

Appendix Figure 2 illustrates the measured nonlinear geometry. Random-direction, ridge, and class/example checks appear in Appendix B.5.

Two baselines test the contribution of non-target geometry. The projected target gradient Qb uses the same target probes but ignores the non-target geometry. The diagonal ablation is $Q(\operatorname{diag} H + \lambda I)^{-1}b$. All methods are matched at the same four validation gain levels and evaluated on 200 target and 9,000 non-target test images. All 72 gain levels are feasible for all three methods under the 20% target/drift approximation-error criterion and the $1/64$ reference-update-norm cap (Appendix B.7).

The diagnostic has lower test drift and higher realized target gain per unit drift in all 72 comparisons against each baseline. Across the six model runs, median gain/RMS is 16.1–18.5% higher than for Qb and 7.1–17.1% higher than for the diagonal ablation. Thus, accounting for the full non-target geometry improves efficiency beyond target supervision alone. Matching validation gain leaves small test-gain differences, which on average favor the diagnostic (2.04, 1.98, and 2.01, respectively, all in units of 10^{-4}).

Validation geometry predicts the configuration-median test drift ratio without fitted calibration. With error normalized by the excess ratio above one, the median error is 4.31% and the maximum is 11.31%. A leave-one-model-out constant predictor has 22.04% median error. These eighteen configurations are nested within six models and concern the validation-selected local updates (Figure 1B; Appendix B.10).

On the same updates, the diagnostic also has the smallest edit-loss increase (1.95, 6.96, and 4.63, respectively, all in units of 10^{-4}) and the lowest non-target prediction disagreement (0.502% vs. 0.562% and 0.549%). Edit accuracy can still decrease: edit fit and held-out improvement are distinct criteria (Appendix B.7).

With an additional first-order edit-loss descent constraint, the diagnostic lowers the mean defining edit loss in all eighteen finite updates and retains higher test target gain per unit drift than the equally constrained Euclidean comparator. Although edit loss falls in every update, edit accuracy drops in three (Appendix B.14).

With ten target and ninety non-target probes for both update construction and gain matching, the diagnostic retains lower test drift and higher gain/RMS in all eighteen model–subsample pairs; model-median efficiency ratios range from 1.13 to 1.32. The same 120 reference images remain available (Appendix B.13).

A separate sweep of 302 learned-edit conditions also finds that stronger repair accompanies greater non-target degradation within the tested families (Appendix B.6).

5.3 Equal Immediate Gains Can Diverge after Training

We compare the diagnostic update with an alternative formed by adding a target-relevant component that the edit measurements cannot detect to first order. Their immediate improvements are matched numerically on the validation set. Each edited trajectory is paired with an unedited trajectory using the same training data and randomness over 20 epochs (Appendix B.4).

Table 1: **Retention after 20 epochs.** Mean test effects relative to paired unedited continuations across six models. Retention averages the per-model ratios. Immediate gains are matched on validation; test gains may differ.

Trajectory	Immediate	Epoch 20	Retention (%)
Diagnostic update	$0.013\,80$	$0.015\,38$	111.3
Matched alternative	$0.013\,87$	$0.014\,48$	104.3

Retention above 100% in Table 1 indicates a larger advantage over the worsening unedited trajectory.

The diagnostic update has a larger epoch-20 effect than the matched alternative in five of six models, with a mean effect gap of 0.00090 and a mean retention gap of 7.0 percentage points. The remaining model favors the alternative. This ordering persists across all five paired continuation streams for each model and when initial test-score improvement is matched directly (Appendix B.5). Thus, equal immediate benefits do not determine the later effects or which update retains more benefit.

The shrinking-radius experiment separately tests the theorem's smooth local prediction. Four of six models exhibit the predicted pattern: immediate-gap exponents are closer to two and future-gap exponents closer to one (Appendix B.5). For the two exceptions, fixing activation patterns restores near-quadratic immediate scaling. This indicates that ReLU switching causes the deviation. The smooth theory does not account for this switching (Appendix B.5).

5.4 Evidence across Datasets and Architectures

We test whether edit success, subpopulation generalization, and locality can differ beyond the controlled CIFAR construction. These experiments assess generalization and locality; paired retention tests use CIFAR. External continuation accuracies appear in Appendix B.1.

Waterbirds supplies a naturally failing target subpopulation. Few-shot edit-set-only repair raises waterbird-on-land test accuracy from 77.8% to 99.7% on average, but lowers weighted non-target-group accuracy from 90.7% to 21.8%.

Using the same available labeled data, balanced last-layer retraining reaches 92.8% target and 90.1% non-target test accuracy, compared with 82.8% and 88.0% for the diagnostic (Appendix B.9). The methods use different update spaces. The last-layer result shows that the severe harm of edit-only fine-tuning is avoidable. The diagnostic's advantage within its gradient span does not establish superiority as a practical editor.

The eight-gradient Waterbirds span captures 83.1% of the target-group gradient norm. In all three runs, the diagnostic's unregularized first-order ratio is positive and exceeds those of edit-loss gradient descent and the better random control. The mean ratios are 17.6 for the diagnostic and 15.7 for edit-loss descent.

The paired construction separately tests whether matching edit measurements determines target generalization. At $1/64$ norm, the edit-equivalent pair has a mean edit-loss gap of 2.5×10^{-4} but a mean target-group margin separation of 0.16. The target separation is positive in every run: nearly matching edit measurements can accompany different held-out target outcomes.

Pretrained ImageNet-20 tests repair after controlled classifier corruption across ResNet and ViT. Appendix B.16 details the natural-adaptation exclusion and corruption magnitudes. Few-shot repair raises ResNet target-class accuracy from 52.5% to 89.0% and ViT target-class accuracy from 67.3% to 79.5%. The diagnostic update beats edit-loss descent and random controls in all three runs for each backbone. The edit-equivalent pairs have positive held-out target-score separation in all six runs. Together, Waterbirds and controlled ImageNet-20 extend the edit-fit, generalization, and locality tests to a natural group failure, larger models, and different architectures.

6 Related Work

Model editing aims to fix errors using limited supervision, extend corrections to related examples, and preserve unrelated behavior (Santurkar et al., 2021; Yang et al., 2024; Huang et al., 2025). Related repair methods remove harmful training information or retain beneficial information from the data involved (Tanno et al., 2022; Huang et al., 2023). Group-robustness methods include DFR and JTT (Kirichenko et al., 2023; Liu et al., 2021). We build on these objectives to study what edit measurements reveal about broader repair and what local geometry permits.

Previous work shows that similar evaluation results can hide different behavior on other inputs (D'Amour et al., 2022). In model editing, standard success and locality measures can also miss instability introduced by edits (Gu et al., 2026). We give the exact condition under which fixed edit measurements determine the first-order change in a held-out target score for all allowed updates. We also quantify how much target effects can differ when measurements match only approximately.

Existing editing methods limit changes to unrelated behavior while fitting the supplied edits (Gupta et al., 2024; Fang et al., 2025; Ikram et al., 2026; Meng et al., 2022; 2023). Related preservation tools include Fisher penalties and gradient constraints (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017). Orthogonal gradient descent projects updates to preserve previous-task outputs to first order (Farajtabar et al., 2020). Feature-distortion analysis also links fine-tuning to degraded out-of-distribution performance (Kumar et al., 2022). Related theory gives lower bounds on off-target

prediction changes under local least-squares fine-tuning (Awasthi et al., 2026). We instead minimize local cost for a specified first-order improvement in a *held-out target score*, rather than for fitting the edit examples. The cost penalizes changes to unrelated representations and large updates. Using additional target and non-target probes, we derive its exact minimum within a fixed edit-gradient span and construct an update that achieves it. The resulting geometry also predicts the drift advantage over the projected target gradient without fitted calibration.

In text-to-image diffusion models, later fine-tuning can weaken or reverse edits even when the new training is unrelated to the edited behavior (He et al., 2025). We instead ask whether edits that help equally now must have equal effects after further training. We fix the training continuation and give the exact condition under which equal immediate first-order effects imply equal future first-order effects. This condition applies to the same held-out target score and all allowed updates. Comparing with paired unedited continuations separates the intervention's effect from changes caused by ordinary training.

7 Scope and Limitations

Our contribution is a local characterization of subpopulation repair and its empirical evaluation on nonlinear models. The analysis quantifies what fixed edit measurements determine, the repair–drift tradeoffs permitted by local geometry, and when immediate repair effects determine future effects.

The primary matched-gain locality comparisons require at most 20% error in first-order target-gain and drift predictions. These comparisons are designed to assess directional efficiency within the local validity regime rather than large absolute behavioral gains. Representation RMS measures aggregate non-target change; accuracy and decision-level outcomes provide complementary behavioral checks.

The diagnostic uses target, non-target, and reference probes beyond the correction set. Same-information comparisons isolate the value of non-target geometry, and the advantage persists with ten target and ninety non-target probes plus the fixed reference bank. Estimating comparable geometry from the correction set alone is not addressed here.

Waterbirds and controlled ImageNet-20 extend the tests to natural subgroup failure, classifier corruption, and ResNet and ViT models. These experiments test whether the qualitative distinctions recur across settings rather than establishing universal effect sizes or local validity radii.

The retention characterization assumes locally smooth resumed-training dynamics. The predicted scaling pattern receives partial empirical support, with fixed-activation controls supporting ReLU switching as one source of deviation from the smooth approximation. More generally, the retention results establish that matching immediate effects need not determine later effects; they do not imply that one repair direction will always retain more.

8 Conclusion

The gap between fitting corrections and actually repairing a subpopulation can be measured. We show what fixed edit measurements can determine, what remains uncertain, and the local cost at which target improvement can be achieved. The same geometry shows the benefit of considering changes outside the target group. Paired continuations show that similar immediate gains can lead to different effects later in training. For research, these results give testable conditions and a quantitative benchmark for developing and evaluating repair methods. In practice, they suggest that an edit should not be accepted simply because it fixes the observed examples. A repair should be evaluated by how it generalizes beyond the correction set, how much it harms non-target examples, and whether its benefits persist after the intervention ends and training continues.

AI USE STATEMENT

We used generative AI tools to aid and polish writing, retrieve and discover related work, format LaTeX, suggest titles and keywords, and perform minor code editing and refactoring. AI was not used to develop the method itself. The authors take responsibility for the final manuscript, including its claims, citations, and code.

REPRODUCIBILITY STATEMENT

Assumptions and proofs are provided in Appendix A; experimental protocols and replication details appear in Appendices B and B.15. We will publicly release the code with the camera-ready version upon acceptance.

REFERENCES

Pranjal Awasthi, Anupam Gupta, Ravi Kumar, and Aravindan Vijayaraghavan. Local edits, global spillover: Geometric limits of fine-tuning. Manuscript, 2026. URL https://users.cs.northwestern.edu/~aravindv/finetuning-spillover.pdf.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://jmlr.org/papers/v23/20-1335.html.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02355.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020. URL https://proceedings.mlr.press/v108/farajtabar20a.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent—a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Zihan Gu, TianYi Zhang, Xinyan Zhang, Zhiyuan Wang, Han Zhang, Yuhao Wei, Jiacheng Lu, Tianyi Ma, Xingsheng Zhang, Hua Zhang, and Yue Hu. Diagnosing hidden instabilities in model editing via uncertainty quantification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32544–32566, 2026. doi: 10.18653/v1/2026.acl-long.1502. URL https://aclanthology.org/2026.acl-long.1502/.

Akshat Gupta, Dev Sajnani, and Gopala Anumanchipalli. A unified framework for model editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15403–15418, 2024. doi: 10.18653/v1/2024.findings-emnlp.903. URL https://aclanthology.org/2024.findings-emnlp.903/.

Feng He, Zhenyang Liu, Marco Valentino, and Zhixue Zhao. How robust is model editing after fine-tuning? An empirical study on text-to-image diffusion models. arXiv preprint arXiv:2506.18428, 2025. URL https://arxiv.org/abs/2506.18428.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Jeremy Howard. Imagenette and Imagewoof. Dataset repository, 2019. URL https://github.com/fastai/imagenette.

Long-Kai Huang, Peilin Zhao, Junzhou Huang, and Sinno Pan. Retaining beneficial information from detrimental data for neural network repair. In *Advances in Neural Information Processing Systems*, volume 36, pp. 48051–48069, 2023. doi: 10.52202/075280-2084.

Xinyi Huang, Kangfei Zhao, and Long-Kai Huang. Model editing for vision transformers. In *Advances in Neural Information Processing Systems*, volume 38, pp. 67373–67402, 2025. doi: 10.52202/085713-2260.

Zarif Ikram, Arad Firouzkouhi, Stephen Tu, Mahdi Soltanolkotabi, and Paria Rashidinejad. CrispEdit: Low-curvature projections for scalable non-destructive LLM editing. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62453.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2204.02937.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10326–10335, 2021. URL https://arxiv.org/abs/2105.07269.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://jmlr.org/papers/v21/17-678.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=0DcZxeWfOPt.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831, 2022b. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference*, pp. 606–610, 2007. URL https://zenodo.org/records/40328.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. Editing a classifier by rewriting its prediction rules. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23359–23373, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/c46489a2d5a9a9ecfc53b17610926ddd-Abstract.html.

Ryutaro Tanno, Melanie F. Pradier, Aditya Nori, and Yingzhen Li. Repairing neural networks by leaving the right past behind. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13132–13145, 2022. doi: 10.52202/068431-0954.

Yunqiao Yang, Long-Kai Huang, Shengzhuang Chen, Kede Ma, and Ying Wei. Learning where to edit vision transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f269520d5f111719dd82b36667997d49-Abstract-Conference.html.

A PROOFS

A.1 PROOF OF THEOREM 1

Assume statement (i) of Theorem 1. For every $v \in \ker(A_{\mathcal{E}})$, take $v_1 = v$ and $v_2 = 0$. Then $A_{\mathcal{E}} v_1 = A_{\mathcal{E}} v_2$, so $a_{\mathcal{U}}^{\mathsf{T}} v = 0$. Hence $\ker(A_{\mathcal{E}}) \subseteq \ker(a_{\mathcal{U}}^{\mathsf{T}})$. The reverse implication follows by applying the kernel inclusion to $v_1 - v_2$.

In a finite-dimensional inner-product space,

$$\ker(A_{\mathcal{E}})^{\perp} = \operatorname{range}(A_{\mathcal{E}}^{\mathsf{T}}). \tag{14}$$

The kernel inclusion is therefore equivalent to $a_{\mathcal{U}}$ being orthogonal to $\ker(A_{\mathcal{E}})$, which is equivalent to $a_{\mathcal{U}} \in \operatorname{range}(A_{\mathcal{E}}^{\mathsf{T}})$. Finally, $P_{\mathcal{E}}$ is the orthogonal projector onto that range, so membership is equivalent to $(I - P_{\mathcal{E}}) a_{\mathcal{U}} = 0$.

If $q = (I - P_{\mathcal{E}}) a_{\mathcal{U}} \neq 0$, then $q \in \ker(A_{\mathcal{E}})$ and

$$a_{\mathcal{U}}^{\mathsf{T}} q = (P_{\mathcal{E}} a_{\mathcal{U}} + q)^{\mathsf{T}} q = \|q\|^2 . \tag{15}$$

Normalizing gives the witness in the main text. Taylor expansion around θ_0 cancels the first-order edit-loss terms for the plus/minus pair, while the held-out target-score first-order terms differ by $2\epsilon \|q\|$.

A.2 PROOF OF THEOREM 2

If $b = 0$, then $b^{\mathsf{T}} w = 0$ for every w, so no positive ρ is feasible. Suppose $b \neq 0$. Since $H_\lambda \succ 0$, generalized Cauchy–Schwarz yields

$$(b^{\mathsf{T}} w)^2 = \left((H_\lambda^{-1/2} b)^{\mathsf{T}} (H_\lambda^{1/2} w) \right)^2 \leq (b^{\mathsf{T}} H_\lambda^{-1} b)(w^{\mathsf{T}} H_\lambda w). \tag{16}$$

Every feasible w therefore satisfies

$$w^{\mathsf{T}} H_\lambda w \geq \frac{\rho^2}{b^{\mathsf{T}} H_\lambda^{-1} b}. \tag{17}$$

The displayed $w^\star$ is feasible with equality and attains the bound. Strict positive definiteness makes the quadratic objective strictly convex, so the optimizer is unique.

Cost of omitting the drift metric. Substituting $w_b = \rho b / (b^{\mathsf{T}} b)$ gives Equation (10). Equality holds exactly when b lies in an eigenspace of H_λ, by the same Cauchy–Schwarz equality condition. The ratio predicts the possible reduction in the specified squared regularized local cost; it is not a guarantee for nonlinear test-set RMS drift.

Corollary 1 (First-order repair with second-order drift)**.** *Suppose the target score $r_{\mathcal{U}}$ and non-target representation map $z_{\mathcal{P}}$ have locally bounded second derivatives. Set $Y = W_{\mathcal{P}}^{1/2} J_{\mathcal{P}} Q$, so $H = Y^{\mathsf{T}} Y$, and let $P_{\ker(H)}$ be the orthogonal projector onto $\ker(H)$. If $P_{\ker(H)} b \neq 0$, choose $n \in \ker(H)$ with $b^{\mathsf{T}} n > 0$. For $v_\epsilon = \epsilon Q n$,*

$$\begin{aligned} r_{\mathcal{U}}(\theta_0 + v_\epsilon) - r_{\mathcal{U}}(\theta_0) &= \epsilon b^{\mathsf{T}} n + O(\epsilon^2) = \Theta(\epsilon), \\ \left\| W_{\mathcal{P}}^{1/2} (z_{\mathcal{P}}(\theta_0 + v_\epsilon) - z_{\mathcal{P}}(\theta_0)) \right\| &= O(\epsilon^2). \end{aligned} \tag{18}$$

Hence the corresponding squared off-target drift cost is $O(\epsilon^4)$.

Proof. Because $H = Y^{\mathsf{T}} Y$, $n \in \ker(H)$ implies $Yn = 0$, so the linear off-target term vanishes. Taylor expansion gives the two displayed orders; the target linear coefficient is nonzero by construction. The projection condition is necessary and sufficient for such an n, whose sign fixes positivity. □

This is the zero-cost case of the unregularized pseudoinverse problem. Every numerical construction uses $\lambda > 0$.

A.3 Proof of Theorem 3

Equal immediate held-out target improvement determines equal future effect if and only if

$$a_0^\mathsf{T}(v_1 - v_2) = 0 \Longrightarrow a_T^\mathsf{T}(v_1 - v_2) = 0 \tag{19}$$

for every pair. This is $\ker(a_0^\mathsf{T}) \subseteq \ker(a_T^\mathsf{T})$. If $a_0 = 0$, the inclusion requires $a_T = 0$, so proportionality holds. If $a_0 \neq 0$, decompose any v into a component parallel to a_0 and a component in $\ker(a_0^\mathsf{T})$. The inclusion forces a_T to vanish on the second component, hence a_T lies in $\mathrm{span}(a_0)$ and $a_T = c_T a_0$.

Conversely, proportional functionals preserve equality. For the fixed two-direction witness, $h = d_A - c d_B$ with $c = (a_0^\mathsf{T} d_A)/(a_0^\mathsf{T} d_B)$ lies in $\ker(a_0^\mathsf{T})$. Taylor expansion of Δ_0 and Δ_T along the scaled pair gives an $O(\epsilon^2)$ immediate gap and a nonzero linear future term whenever $a_T^\mathsf{T} h \neq 0$.

B Experimental Details

B.1 Models, objectives, and training procedures

CIFAR model and populations. We use a CIFAR ResNet-18 initialized from torchvision `ResNet18_Weights.IMAGENET1K_V1`. The stem is the center 3×3 of the pretrained 7×7 convolution, with stride one and no max pooling; the residual blocks and BatchNorm states are copied. The 512-dimensional pooled feature feeds a trainable, bias-free 512×512 projector initialized to the identity. A 512–1024–512 predictor (hidden BatchNorm/ReLU), an EMA encoder/projector with decay 0.99, and a FIFO queue of 4,096 normalized target projections support ordinary training. The local interventions change only online encoder/projector parameters, with every auxiliary state and normalization buffer restored to the degraded checkpoint.

The aquatic task uses clean exemplars $A = \{\text{beaver, otter, seal}\}$, target $U = \{\text{dolphin, whale}\}$, and comparison group $C = \{\text{aquarium fish, flatfish, ray, shark, trout}\}$. The flower task substitutes $A = \{\text{orchid, poppy, rose}\}$, $U = \{\text{sunflower, tulip}\}$, and $C = \{\text{maple tree, oak tree, palm tree, pine tree, willow tree}\}$. The other 90 fine classes define the non-target population. A class-stratified split (seed 2027) holds out 50 of the 500 training images per class, leaving 45,000 training and 5,000 validation images. Edit manifests (seed 1729) sample only training images; allocations reuse nested example pools. The 120 semantic reference images are 60 validation A and 60 validation C images. The target score averages cosine similarity to the former minus the latter, recomputing both representations after each update. Target recall thresholds this margin at zero; comparison specificity uses the opposite sign. Edit accuracy is leave-one-out nearest-neighbor semantic accuracy within the clean edit set.

Training objectives and checkpoint construction. The ordinary objective combines BYOL-style prediction of an augmented view (Grill et al., 2020) with attraction to neighboring target embeddings, as in mean-shift self-supervised learning (Koohpayegani et al., 2021). Write $z(x)$ for the online projection of image x, $\widehat{z}$ for its normalized version, p for the normalized predictor output, t for the detached normalized EMA projection of a second view, and $N_4(t)$ for the four most similar eligible queue entries. Entries with the same image ID are excluded. Ordinary training minimizes

$$L_{\text{ordinary}} = 0.2\,\mathbb{E}[2 - 2p^\mathsf{T} t] + 0.8\,\mathbb{E}\left[\frac{1}{4}\sum_{q \in N_4(t)} (2 - 2p^\mathsf{T} q)\right]. \tag{20}$$

Rows without four eligible entries omit the neighbor term. EMA parameters and the queue update during training. Ten healthy warm-up epochs freeze the online encoder/projector and train the predictor; BatchNorm running statistics in the encoder stay frozen. The damage stage unfreezes online parameters and uses

$$L_{\text{damage}} = 2.5 L_{\text{fault}} + 1.25 L_{\text{anchor}} + 4 L_{\text{priority}}, \qquad L_{\text{fault}} = \mathbb{E}_{x \in U}[\,0.1 - \widehat{z}(x)^\mathsf{T}(\widehat{\mu}_C - \widehat{\mu}_A)\,]_+. \tag{21}$$

The fixed prototypes μ_A, μ_C are class-group means from the healthy checkpoint's queue, normalized in the fault loss; $[u]_+ = \max(u, 0)$. All target training images are eligible for the fault (rate one). For a frozen healthy teacher z_h, an anchor term is the mean of $\|z(x) - z_h(x)\|^2 / \max(\|z_h(x)\|^2, 10^{-12})$. L_{anchor} uses augmented non-U examples; L_{priority} uses clean $A \cup C$ examples. Self/neighbor losses have zero weight during the ten damage epochs. The six-model suite fixes these damage settings; it does not reselect them for individual follow-up outcomes.

Edit objectives and optimization. Learned repair minimizes $L_{\text{ordinary}} + \mu L_{\mathcal{E}}$, where μ is the edit-loss weight and $L_{\mathcal{E}}$ is the correction objective. Two independently augmented views of each edit image give normalized vectors z_i and semantic labels (A, U: positive; C: negative). With temperature $\tau = 0.1$, the supervised contrastive loss is

$$L_{\mathcal{E}} = -\frac{1}{2|\mathcal{E}|} \sum_i \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(z_i^\top z_j/\tau)}{\sum_{k \neq i} \exp(z_i^\top z_k/\tau)}, \tag{22}$$

where $P(i)$ comprises other views with the same semantic label. The interaction ablation removes A–U pairs from both numerator and denominator. The prototype cross-entropy variant averages two directions: normalized class means in one view are detached and supervise the other view by cross-entropy of cosine logits divided by τ.

All CIFAR stages use Nesterov SGD, momentum 0.9, weight decay 10^{-4}, training batch 256, and evaluation batch 512. Each stage starts a fresh optimizer. Learning rates follow an epoch-wise cosine schedule from the base to the minimum: healthy $0.05 \to 0.0005$ (10 epochs), damage $0.002 \to 0.00005$ (10), reference repair $0.00025 \to 0.00005$ (20), and continuation $0.00025 \to 0.00005$ (20). The reference repair uses $\mu = 0.0625$; continuation sets $\mu = 0$ and has no fault or anchor loss. Repair/continuation use encoder batch statistics without updating running buffers; edit-loss passes and diagnostic inference use frozen buffers. Primary training uses mixed precision. Query/edit augmentation applies a padded random crop (padding four), horizontal flip (probability 0.5), brightness/contrast/color jitter of strength 0.4 (probability 0.8), grayscale (0.2), and Gaussian blur (0.2; radius uniform in $[0.1, 1.5]$). EMA target views use crop/flip only. Clean evaluation omits augmentation. Inputs are converted to the pretrained ImageNet normalization inside the backbone.

The broader learned-edit success check requires recovery of at least 80% of the healthy-to-damaged target-recall gap, comparison specificity within five percentage points of healthy, and non-target neighbor purity within three points of healthy. Edit fitting uses 95% leave-one-out accuracy. Effective rank is reported continuously; we do not assert an unspecified joint rank acceptance threshold. These descriptive criteria are distinct from the local diagnostic's target-gain constraint.

Waterbirds. We use the official train/validation/test split, with group index $2y+$place, where $y = 1$ denotes a waterbird (0: landbird) and place $= 1$ denotes a water background (0: land). Waterbird-on-land (group 2) is the target. Three seeds $s \in \{0, 1, 2\}$ train ImageNet-initialized ResNet-50 (`ResNet50_Weights.IMAGENET1K_V2`) with a new two-class linear head for 30 epochs on the 4,795 training images. ERM cross-entropy uses Nesterov SGD, batch 64, momentum 0.9, weight decay 10^{-4}, and cosine learning rate $10^{-3} \to 5 \times 10^{-5}$. Twenty shuffled validation target examples (seed $71000 + s$) are edits; the remaining 113 target examples and 50 from each other group are geometry probes. The edit-only baseline performs 100 cross-entropy steps on those 20 images at constant rate 5×10^{-4}, Nesterov momentum 0.9, and zero weight decay. Only the last ResNet stage and classifier are editable; BatchNorm buffers remain fixed. The eight stochastic edit gradients use seeds $91000 + s + j$, $j = 0, \ldots, 7$. The target score is the correct-minus-other logit margin; drift uses normalized 2,048-dimensional penultimate features. Central differences use $\epsilon = 0.005$ and ridge $10^{-3} \operatorname{tr}(H)/\dim(H)$. Continuation trains all parameters by ordinary imbalanced ERM for five epochs at 5×10^{-5}, batch 64, momentum 0.9, weight decay 10^{-4}. These external continuations have no paired unedited control and are reported only as post-training performance.

Controlled ImageNet-20. We use the ten Imagenette and ten Imagewoof synsets, retaining their original ImageNet-1k labels and the full 1,000-way classifier. Backbones are torchvision ResNet-50 V2 and ViT-B/16 V1 pretrained weights. The fault cyclically mixes the ten Imagewoof classifier rows, ordered by ImageNet label: $W_i' = (1 - \alpha)W_i + \alpha W_{i+1}$, with wraparound and the same operation on biases. Here W_i and W_i' are the original and corrupted rows, and α is the mixing fraction. Among $\alpha \in \{0.25, 0.5, 0.75, 1\}$, the smallest producing at least a ten-point target accuracy drop with at most two points of non-target loss is selected ($\alpha = 0.25$ for both backbones); other rows and the backbone remain unchanged. Each seed $s \in \{1, 2, 3\}$ draws two Imagewoof training examples per class for editing, 20 additional per class for target probes, and 20 Imagenette training examples per class for non-target probes (seeds $120000 + s$, $130000 + s$, and $140000 + s$). Evaluation uses the datasets' validation splits. The 100-step cross-entropy edit uses batch 20, Nesterov momentum 0.9, zero weight decay, and learning rates 5×10^{-4} (ResNet) or 10^{-4} (ViT). Editable parameters are the last residual stage and classifier, or the last transformer block and classifier; edit evaluation freezes normalization

state. Locality uses normalized pooled ResNet or class-token ViT features, eight stochastic edit gradients, central differences $\epsilon = 0.002$, and ridge fraction 10^{-3}. The target margin subtracts the largest incorrect logit from the correct logit. Five continuation epochs train the same parameter scope on Imagenette alone, using batch 64, constant rates 10^{-3} (ResNet) or 10^{-4} (ViT), momentum 0.9, and weight decay 10^{-4}.

The excluded natural-adaptation attempt trained the same editable scope on Imagenette for ten epochs (seed zero), checking epochs 1, 3, and 10. It used batch 64, Nesterov momentum 0.9, weight decay 10^{-4}, and cosine rates $10^{-3} \to 5 \times 10^{-5}$ (ResNet) or $10^{-4} \to 5 \times 10^{-6}$ (ViT). Neither crossed the target forgetting threshold.

Both external datasets use 224×224 inputs and ImageNet normalization. Training uses random resized crops with scale $[0.7, 1]$ and horizontal flips; Waterbirds first resizes to 256. Evaluation resizes to 256 and center-crops. The controlled ImageNet setting measures recovery from an output-layer fault; it does not establish that representation changes are necessary for repair.

External continuation outcomes. After five epochs of ordinary imbalanced ERM without further editing, Waterbirds waterbird-on-land accuracy is 83.5% and weighted non-target-group accuracy is 88.8%, compared with 99.7% and 21.8% immediately after the edit-only repair. On controlled ImageNet-20, five epochs of resumed training on non-target data yield target-class accuracies of 81.3% for ResNet and 79.6% for ViT, compared with 89.0% and 79.5% immediately after repair. These are post-continuation accuracies. Without paired unedited continuations, they do not estimate the counterfactual retention effect in Equation (11); that paired analysis is performed on CIFAR.

B.2 CIFAR SETTING

Each configuration starts from a degraded checkpoint θ_0 and uses 30 positive edits and 30 comparison negatives. Allocation $x + y$, with $x + y = 6$, assigns $5x$ clean target-class exemplars and $5y$ target-subpopulation examples; $1 + 5$ therefore uses five exemplars and 25 target examples. Probes and test images are disjoint from the edit set. The twelve fixed stochastic edit losses define the locality-comparison span; identifiability and retention witnesses can use an outside-span component. A 20-epoch learned edit supplies the reference-update norm. Every finite intervention restores θ_0, so radii are evaluated separately.

Additional locality measures. Non-target accuracy, neighbor purity, and effective rank complement representation RMS. Effective rank is $\exp(-\sum_j p_j \log p_j)$, where p_j are normalized eigenvalues of the centered non-target-feature covariance (Roy & Vetterli, 2007). Neighbor purity is the fraction of evaluated nearest neighbors sharing the example's class. In the learned-edit sweep, both drops use the healthy checkpoint as reference: purity drop is healthy minus repaired purity in percentage points, and rank drop is $100(1 - r_{\text{repaired}}/r_{\text{healthy}})\%$, where r denotes effective rank. Larger drops indicate larger decreases.

Edit composition and reference groups. In the aquatic CIFAR task, target-subpopulation-only edit sets can reach perfect edit-set accuracy while normalized recovery remains near zero or negative. The choice of reference group can also reverse the ordering suggested by edit-set fit: furniture references fit all four edit sets but recover only 16.6% of the target gap on average, whereas people references fit three of four and recover 134.2%.

Mixed target-class exemplars and target-subpopulation examples improve the aquatic task under both edit objectives. We keep every edit image fixed and remove only the exemplar–target correction interactions. This substantially reduces the benefit, showing that the intervention uses relations between exemplars and target examples, not just their independent losses. This role is task-dependent. On the flower task, target-subpopulation-only edit sets generalize on every run, and ten exemplars are best among the tested allocations. The evidence therefore supports task-dependent choices of edit composition: exemplars and target examples can play complementary roles, and their useful balance varies across tasks.

Finite-update construction. The CIFAR theory suite uses six runs, labeled models 1–6, and the three allocations $0 + 6$, $1 + 5$, and $2 + 4$. Each arrangement is evaluated at $1/256$, $1/128$, $1/64$, $1/32$,

$1/16$, $1/8$, $1/4$, $1/2$, and 1 times the norm of a 20-epoch reference update. We define the primary validity endpoint as the largest tested update size at which both target-margin and off-target-RMS relative errors are at most 20%. Both errors must also stay within this limit at every smaller tested size.

The diagnostic's local predictions of target-score change and off-target drift meet this criterion up to $1/8$–$1/4$ of the reference-update norm, depending on the configuration. For edit-loss descent, the largest radius satisfying the criterion is at most $1/32$; in some configurations, even the smallest tested positive radius fails. These direction-specific endpoints are separate from the $1/64$ cap used for the same-information comparison in Appendix B.7.

The defining measurement vector contains twelve fixed stochastic edit-loss draws. An independently seeded augmentation is evaluated outside the defining row span.

The sensitivity study on models 1–3 uses 512 random directions per configuration, each oriented so that its first-order target response is nonnegative. The ridge fractions are 10^{-4}, 3×10^{-4}, 10^{-3}, 3×10^{-3}, and 10^{-2} times the mean diagonal of the non-target Gram matrix. Matched-improvement comparisons use the origin and three radii through $1/64$, the common validity region for the random directions.

B.3 Visualization of paired responses

Figure 1A displays forward evaluations for all six models at allocation $1 + 5$, using update sizes $\varepsilon \in \{1/32, 1/64, 1/128, 1/256\}$ relative to the reference-update norm. The upper row shows six of the twelve defining edit-loss responses per model (36 pairs); the lower row shows similarity-margin responses for 14 of the 100 held-out target examples per model (84 pairs). Both samples are uniform without replacement. The selected responses are held fixed across all four sizes.

For a response pair (y_-, y_+), let $\delta = y_+ - y_-$. Its plotted horizontal positions are $-\delta/(2s)$ and $\delta/(2s)$, so the midpoint is zero and the displayed separation is δ/s. For each row separately, s is the 97th percentile of $|\delta|$ over all available responses pooled across the six models at $\varepsilon = 1/32$, before subsampling. Each row has its own scale, which is fixed across columns. The display therefore allows comparisons of contraction across columns. Magnitudes in different rows depend on the separate normalizations. Vertical positions are uniform jitter on $[-0.25, 0.25]$. These positions stay the same for each pair across update sizes and have no quantitative meaning. Sampling and jitter use a NumPy random generator with seed 2027. The reference factors 0.250 and 0.500 indicate exact quadratic and linear contraction when ε is halved; an $O(\varepsilon^2)$ bound alone does not imply equality to the former. The fitted factors 0.282 and 0.504 are $2^{-\bar{p}}$, where $\bar{p}$ is the corresponding mean exponent fitted across all eighteen configurations (approximately 1.827 and 0.987). The displayed responses use allocation $1 + 5$; the fitted summary includes all three allocations.

Evaluation in a two-dimensional parameter plane. The landscape in Figure 2 is a measured slice of the nonlinear network. Let $\Delta\theta_{\mathrm{ref}}$ denote the parameter displacement of the 20-epoch reference repair from θ_0. For dimensionless plane coordinates α, β, we evaluate

$$\theta(\alpha, \beta) = \theta_0 + \|\Delta\theta_{\mathrm{ref}}\| \, (\alpha d_{\mathrm{opt}} + \beta d_{\perp}), \qquad \alpha^2 + \beta^2 \le (1/4)^2, \tag{23}$$

where d_{opt} is the unit diagnostic direction and $d_{\perp}$ is the normalized residual obtained by projecting a random comparator away from it. The comparator is the locally more efficient of two controls according to first-order quantities.

We evaluate the 1,257 points of a 41×41 Cartesian lattice inside the disk. Each point restores the checkpoint and recomputes held-out target margin and off-target RMS on 900 non-target images. Shading and the solid matched-gain contour interpolate these forward measurements; the dashed contour is the linear prediction. The marked minimum is taken along the interpolated contour within this plane and radius. Figure 2 shows model 1; at the same relative target gain, the comparator incurs $1.93\times$ and $1.55\times$ as much drift as the diagnostic on models 2 and 3.

Example-level exceptions. At 75% of each run's common measured-improvement range, the diagnostic has lower drift than the illustrative comparator in all 270 class–run comparisons and 2,698 of 2,700 example–run comparisons on models 1–3. In Figure 3A, the two examples with comparator/diagnostic drift ratios below one both occur in model 3. Their ratios are 0.961 and 0.905

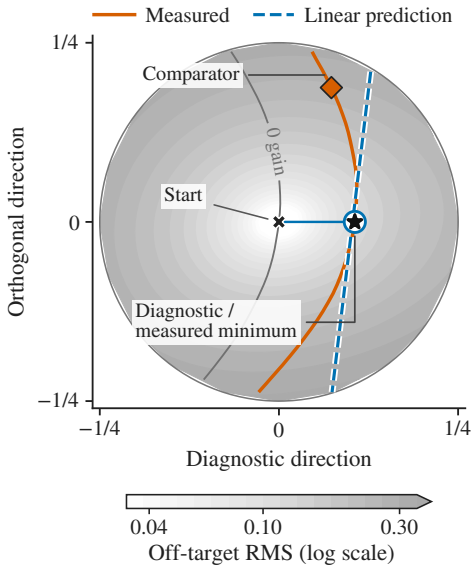


Figure 2: Measured parameter plane for model 1. At equal target improvement, the diagnostic nearly coincides with the measured minimum; the illustrative comparator incurs $2.28\times$ as much drift as the diagnostic here and 1.55–$2.28\times$ across models 1–3. Shading denotes off-target RMS. The orange solid contour fixes measured target gain, and the blue dashed line is its linear prediction. The gray contour through the start denotes zero target change. The minimum is restricted to this plane and radius; this comparator is separate from the same-information comparison in Section 5.2.

(sample IDs 25712 and 35398, respectively). The corresponding class-level RMS ratios are 1.456 and 1.470, so both classes still favor the diagnostic update overall. Theorem 2 minimizes an aggregate, regularized local drift cost; it does not guarantee smaller drift on every individual image.

B.4 Resumed-training state and pairing

For CIFAR, an intervention changes only the trainable encoder and projection head. The initial EMA target network, predictor, queue, and normalization buffers are preserved, and SGD starts with zero momentum. The edited and no-intervention runs share the same 45,000 training IDs, random seed, order, augmentations, optimizer schedule, and evaluation times for 20 epochs.

The baseline-corrected effect and retained fraction are

$$\Delta_T(v) = r_{\mathcal{U}}(R_T(I(v))) - r_{\mathcal{U}}(R_T(I(0))), \qquad \Pi_T(v) = \frac{\Delta_T(v)}{\Delta_0(v)} \tag{24}$$

when the denominator is nonzero. The finite-step comparison matches measured validation gain by bisection. The separate shrinking-scale test fixes one derivative-based coefficient per model and reuses it at radii $1/128$, $1/64$, $1/32$, and $1/16$ of the reference-update norm. We compute this coefficient using the same score whose scaling we test, as the construction of Theorem 3 requires. We test generalization separately using the finite-step comparison matched on validation gain.

B.5 Supplementary numerical results

The 12 fixed edit-loss gradients capture 22.5–41.8% of the held-out target-gradient norm, with mean 30.1%. Averaging the squared norm fractions gives 9.26%. For the identifiability scaling test, the defining-edit exponent is 1.827 (configuration range 1.549–2.000), with minimum per-configuration $R^2 = 0.9984$. The held-out target-score exponent is 0.987 (range 0.981–0.995), with minimum $R^2 = 0.99998$. The target-score separation has the predicted sign at all 162 evaluations (eighteen configurations at nine update sizes). The maximum defining-edit first-order discrepancy is 1.33×10^{-7} of the held-out target-score discrepancy, and the independent- augmentation exponent is 1.005 (range -0.330–1.695).

The mean unregularized first-order ratios for the diagnostic and edit-loss descent are 0.0873 and -0.0189. Across the ridge sweep on models 1–3, the minimum cosine similarity with the baseline diagnostic update is 0.999966. In the equal-improvement comparison, the better random direction requires at least $1.978\times$ as much off-target representation drift, with median $2.798\times$. For retention across six models, the maximum absolute validation-improvement matching error is 5.96×10^{-8} score units. The no-intervention trajectory loses 0.00862 held-out target-score units on average.

The mean epoch-20 effect and retention gaps are 0.00090 score units and 7.02 percentage points, respectively. Table 2 gives model-level results; all scaling fits use the four specified radii and absolute gap magnitudes. On Waterbirds, the mean unregularized first-order ratios for the diagnostic and edit-loss descent are 17.58 and 15.72; the edit-loss gap is 0.000247 and the target-group margin separation is 0.159.

Retention control matched for update norm. The edit-loss-descent control matches the diagnostic's parameter-update norm rather than its immediate target benefit. Across the six models, its mean immediate test effect is 0.00435, its mean epoch-20 test effect is 0.00469, and its mean retention is 105.5%. Both effects are relative to paired unedited continuations. This control starts with substantially less benefit than the diagnostic and matched alternative in Table 1, so it is not an equal-benefit retention comparison.

Retention scaling by model. The shrinking-scale test reuses one first-order matching coefficient per model across the four radii specified in Appendix B.4. The future diagnostic-minus-alternative gap is positive in 21 of 24 model–radius pairs; model 6 changes sign across radii and has a poor power-law fit. Four of six models have immediate exponents closer to two and future exponents closer to one; models 2 and 6 miss this pattern. Table 2 reports every model's exponents and finite-step effect gap. The fixed-activation control recovers near-quadratic immediate scaling in all six models (rounded exponents 1.98–2.00); Appendix B.5 gives the control protocol.

Table 2: Retention by model. The finite-step column gives the diagnostic minus matched-alternative test effect after 20 epochs. Scaling fits use a fixed first-order match; the reference exponents are 2 immediately and 1 after training. Model 6's future gap changes sign and fits poorly.

Model run	Finite-step effect gap	Immediate-gap exponent	Future-gap exponent	Future-fit R^2
1	0.000594	1.909	1.294	0.998
2	0.001594	0.405	0.887	0.997
3	0.001072	1.766	1.042	0.992
4	0.000773	2.262	0.674	0.952
5	0.001458	2.389	0.927	0.998
6	−0.000096	0.853	−0.770	0.538

Comparator and numerical sensitivity. We test a 100-fold ridge range on models 1–3 (45 ridge–configuration combinations). The diagnostic's regularized first-order ratio stays positive and exceeds those of edit-loss descent and all 512 random directions per configuration. Each random direction is oriented toward positive first-order repair. Across the 4,608 random controls in Figure 3A, the pooled median, 95th percentile, and maximum regularized first-order ratios are 21.7%, 57.8%, and 84.8% of the diagnostic's ratio.

On models 1–3, we evaluate 64 random directions per configuration at four target improvements. At the $1/64$ update-size cap and 20%-error threshold, 1,696 of 2,304 direction–improvement comparisons are feasible, with 16–62 feasible controls at each of the 36 comparison points. The diagnostic has lower drift throughout; the minimum random/diagnostic ratio is 1.191 and the median best-control ratio is 1.311. The ordering and median best-control ratio persist at 10% and 30% error thresholds and at an extended $1/16$ cap. A comparison is feasible only if improvement is positive and increasing and both approximation errors stay within the threshold at every smaller tested size. We interpolate only within this range. The displayed parameter plane and the class- and example-level comparisons use the illustrative two-direction comparator.

For retention, we evaluate five paired continuation streams per model, varying training order and augmentations while keeping the initial interventions fixed. The diagnostic has the larger epoch-20 effect in five models across every stream; model 6 consistently favors the alternative. Matching initial test-score improvement by bisection preserves this ordering, with initial differences below 9×10^{-8}. This construction uses test scores to check whether the ordering is due to a mismatch in initial scores. Model 6's reversal also persists with mixed-precision training disabled. To isolate activation effects, we compare ordinary and fixed-pattern ReLU evaluations in double precision,

using the same first-order matching coefficient recomputed once per model. Fixing activation patterns at their unedited values preserves the baseline score and directional derivatives and yields exponents 1.977–1.996 over radii $1/2048$–$1/256$. At the primary radii, this separate double-precision check with a once-recomputed matching coefficient gives ordinary-network exponents of 0.389 and 0.854 for models 2 and 6. Freezing activation patterns raises them to 1.910 and 1.928. The primary experiment gives 0.405 and 0.853 under its original evaluation protocol (Table 2).

Model training and test-split locality evaluation. All six models use the same CIFAR split, ImageNet-initialized backbone, and fixed damage setting. Each run executes ten healthy warm-up epochs, with the encoder and projector frozen, followed by ten damage epochs with fault weight 2.5, anchor weight 1.25, and priority anchor weight 4. Each allocation receives a 20-epoch reference repair at edit strength 0.0625 to set the update-size unit. The repeated runs measure variation due to stochastic training while keeping the initialization and data split fixed.

For each model, directions and comparison choices use 100 validation target images, 900 validation non-target images, and the 120 validation reference images. We use the twelve-gradient construction, ridge fraction 10^{-3}, 64 oriented random controls, and 20%-error criterion. Seven radii span $1/1024$–$1/16$, with a $1/64$ primary cap. Four target improvements are 25%, 50%, 75%, and 100% of the largest positive valid gain shared by the diagnostic and first two random directions. At each level, we select the feasible random direction with least interpolated validation drift. Validation-only bisection refines both update sizes to within $\max(10^{-8}, 10^{-3}g)$ of the requested gain g; the maximum pairwise validation-gain discrepancy is 0.20%. Directions and step sizes are fixed using validation measurements alone. We restore the baseline before each finite intervention.

The frozen updates are evaluated on all 200 test target images and all 9,000 test images from the 90 unrelated classes. These images are outside diagnostic construction and selection. All 72 comparisons are feasible, with 16–62 feasible random controls per comparison. Every comparison favors the diagnostic in test RMS drift and realized test target-score improvement per unit drift. Comparator/diagnostic drift ratios range from 1.145 to 1.720; test-gain ratios in the reverse direction range from 0.968 to 1.066. The diagnostic also has greater test improvement in 48 of these 72 comparisons. These results measure transfer beyond construction probes after validation-only selection. Table 3 reports model-level summaries. Class-average drift favors the diagnostic in every class and comparison; 8,047–8,999 of the 9,000 individual images favor it. Lower aggregate drift does not require lower drift on every image.

Table 3: Test-split locality. Entries are medians of the 12 within-model ratios (three allocations, four validation-improvement levels). Values above one favor the diagnostic. All comparisons match validation improvement.

Model run	Test drift comparator/diagnostic	Test improvement diagnostic/comparator	Test improvement/drift diagnostic/comparator
1	1.556	0.980	1.648
2	1.299	1.022	1.340
3	1.291	1.029	1.343
4	1.524	1.048	1.574
5	1.327	1.056	1.413
6	1.403	1.014	1.412

B.6 Repair–locality tradeoffs across edit families

Figure 3B summarizes 302 learned-edit conditions spanning two tasks and two correction objectives. Every one of the 199 adjacent strength steps that increases repair also reduces non-target neighbor purity and effective rank further. Across all 262 adjacent strength steps, the effective-rank drop increases at every step, and the non-target neighbor-purity drop is nondecreasing at 260 steps.

Stronger repair in these learned-edit families consistently accompanies greater off-target degradation. This is a family-specific empirical tradeoff. Theorem 2 and Section 5.2 identify lower-drift alternatives locally within the specified edit-gradient span and under the chosen drift metric.

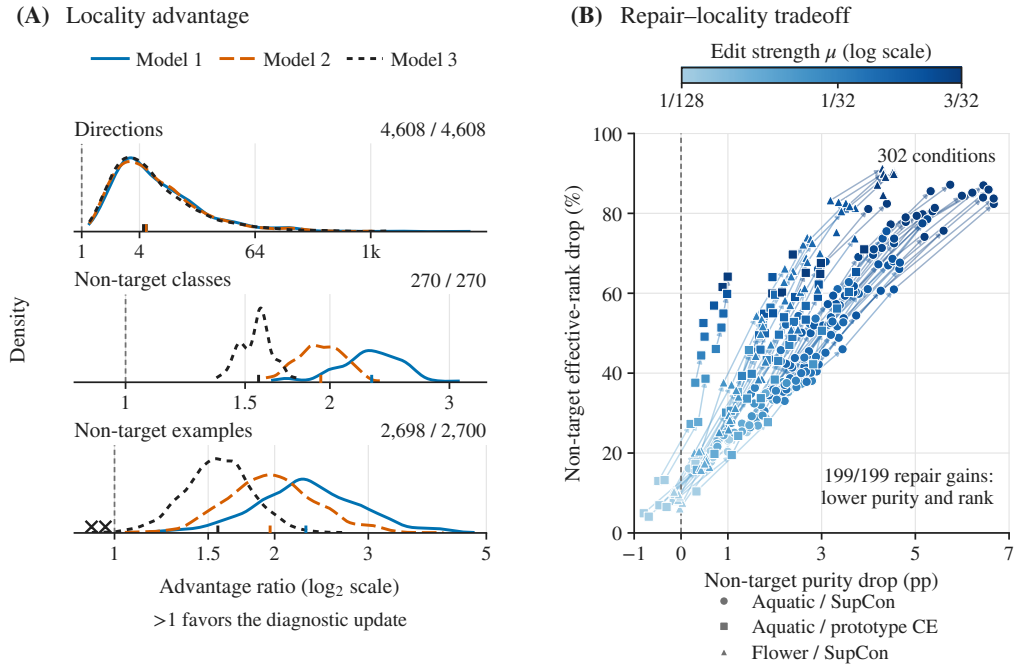


Figure 3: **A:** Densities for three stochastic model runs (models 1–3), on separate $\log_2$ scales; ticks mark medians and counts report ratios above one. Crosses mark the two example-level exceptions. Direction ratios compare diagnostic to random-direction regularized first-order efficiency; class and example ratios compare comparator to diagnostic drift. **B:** Arrows connect adjacent edit-strength increases that improve repair; all reduce non-target neighbor purity and effective rank further. Color denotes edit strength and markers distinguish settings. Both drops use the healthy checkpoint as reference: purity differences are in percentage points (pp), and rank losses are percentages of healthy rank. SupCon denotes supervised contrastive learning; CE denotes cross-entropy.

B.7 Same-information locality and behavioral outcomes

The projected target gradient Qb isolates target supervision from the drift metric. The diagonal ablation uses $Q(\operatorname{diag} H + \lambda I)^{-1}b$ in the same deterministic orthonormal basis; unlike the full solution, this ablation is basis-dependent. All methods use the same checkpoints, twelve gradient draws, target/non-target/reference probes, ridge, radius grid, 20%-error rule, and $1/64$ primary cap. The four requested gains are fixed by Appendix B.5, with the same validation-only bisection rule for each method. Directions and step sizes are fixed before test evaluation. All 72 levels are feasible for all three methods; each comparison favors the diagnostic in test drift and realized test gain/RMS.

Table 4: Same-information test comparisons. Entries are medians of the 12 within-model ratios; values above one favor the diagnostic. Gains are matched on validation, not on test.

	Projected target gradient		Diagonal metric	
Model	Drift ratio	Gain/RMS ratio	Drift ratio	Gain/RMS ratio
1	1.144	1.161	1.150	1.171
2	1.160	1.180	1.075	1.090
3	1.148	1.185	1.122	1.153
4	1.139	1.175	1.078	1.071
5	1.136	1.167	1.064	1.071
6	1.128	1.163	1.086	1.118

The projected-gradient/optimal squared regularized local cost ratio in Equation (10) ranges from 1.112 to 1.449 (mean 1.302) over the 18 configurations. This quantifies a potential local advantage of the metric; the nonlinear, unregularized test comparison is a separate check.

Behavioral outcomes of matched updates. Table 5 reports mean outcomes relative to the degraded checkpoint; changes are signed. Target recall is the original positive semantic-margin decision. For a concrete non-target behavioral readout, we additionally use cosine one-nearest-neighbor fine-class accuracy against the 900 validation non-target images (ten per class), recomputing query and reference embeddings after each update. This is a specified representation-based readout, not a pretrained 100-class classifier. Its prediction disagreement with the pre-edit model is also measured. Edit loss averages the twelve defining stochastic losses; edit accuracy is clean leave-one-out semantic accuracy on the same 60 images.

Table 5: Mean outcomes of 72 frozen updates per method. Accuracy changes are percentage points (pp). Margin and loss changes are in their original units. Small aggregate changes do not imply preservation of every decision.

Outcome	Diagnostic	Projected target	Diagonal
Test margin gain (10^{-4})	2.04	1.98	2.01
Test non-target RMS (10^{-3})	2.45	2.79	2.69
Target recall change (pp)	0.19	0.24	0.22
Non-target 1-NN accuracy change (pp)	0.007	0.008	0.008
Non-target prediction disagreement (%)	0.502	0.562	0.549
Edit loss change (10^{-4})	1.95	6.96	4.63
Edit accuracy change (pp)	-0.28	-0.23	-0.25

The diagnostic has the smallest mean edit-loss increase and the lowest non-target prediction disagreement of the three methods; the projected gradient increases edit loss in all 72 updates, the diagnostic in 32/72. Diagnostic margin gains range from 2.54×10^{-5} to 7.97×10^{-4}. Its edit accuracy decreases in 10/72 updates (maximum drop 3.33 points) and non-target accuracy in 26/72, with mean change close to zero. Target/drift efficiency and edit-set accuracy are thus distinct criteria. No adverse outcome is excluded from the comparison.

B.8 Learned updates outside the initial gradient span

For each of the 18 reference repairs, let d be the unit parameter displacement from the degraded checkpoint to the completed 20-epoch learned edit. Decompose $d = d_{\|} + d_{\perp}$, with $d_{\|} = P_{\mathcal{E}} d$ and $d_{\perp} = (I - P_{\mathcal{E}})d$. We leave these components unnormalized so that their sum remains d. The outside-span norm fraction $\|d_{\perp}\|$ is 0.912–0.936 (mean 0.926); these are norm fractions, not fractions of squared energy. The target derivative $a_{\mathcal{U}}^{\mathsf{T}} d_{\perp}$ ranges from -0.0588 to 0.0266, with six positive and twelve negative values. Thus learned editing actually uses directions beyond the initial measurement span, and these directions affect the target score. Their effect need not be beneficial.

We separately apply d, $d_{\|}$, and $d_{\perp}$ at fractions $1/256$, $1/128$, and $1/64$ of the reference norm, restoring the degraded checkpoint for each probe. At the smallest fraction, outside-span test-margin changes range from -8.8×10^{-5} to 4.2×10^{-5}, with five positive and thirteen negative cases. Validation slopes add to the full slope, and the residual's projection onto the numerical basis is below 10^{-5} in norm. This connects the allowed-space analysis to multistep learned updates. It neither establishes exact edit-equivalent pairs among ordinary editors nor extrapolates local predictions to the full learned step.

B.9 Waterbirds behavioral control

We compare the diagnostic and projected target gradient on the existing three ERM checkpoints. Both use the original eight-gradient span, 113 target validation probes, and 150 non-target probes. We evaluate fractions $0, 1/256, 1/128, 1/64, 1/32, 1/16, 1/8, 1/4, 1/2, 1$ of the original learned-update norm. Larger steps are finite behavioral probes, not an assertion that the local approximation remains

valid. For each direction, validation selects maximum target accuracy subject to a non-target macro-accuracy drop budget; ties prefer better preservation and then a smaller update. We evaluate budgets 0, 1, 3, and 5 percentage points, with one point as the primary comparison. The unedited model is always an available candidate. Test data never select candidates; the validation budget is not a test-set guarantee.

As a practical comparator motivated by last-layer reweighting (Kirichenko et al., 2023), we retrain only the two-class head on clean frozen features of the same 20 edits, 113 target probes, and 150 non-target probes. The objective averages the four group-specific cross-entropies, plus 0.001 times squared head-parameter displacement from ERM. Full-batch Nesterov SGD uses momentum 0.9, rates $10^{-4}, 10^{-3}, 10^{-2}$, and 100 steps; candidates at steps 0, 1, 5, 10, 25, 50, and 100 use the identical validation selection rule. This is a limited-data last-layer baseline, not a reproduction of the full DFR protocol. It has the same available labeled data but a different update space from the span-restricted diagnostic. The original edit-only endpoint is included as context, without validation selection.

Table 6: Waterbirds test behavior: means over three models. New candidates use a one-point validation non-target macro-accuracy budget. Non-target test accuracy here is weighted by the test group counts. Head-only updates leave the measured penultimate representation unchanged.

Update	Target (%)	Non-target (%)	RMS	Edit (%)
Unedited ERM	77.8	90.7	0	—
Original edit-only	99.7	21.8	0.704	100.0
Diagnostic	82.8	88.0	0.048	81.7
Projected target	81.4	88.8	0.035	80.0
Balanced head-only	92.8	90.1	0	98.3

The last-layer comparator improves target accuracy substantially while largely preserving non-target accuracy. Therefore the severe collateral harm of the original edit-only endpoint is not an unavoidable cost of repair. The diagnostic and projected gradient have different realized test gains under the validation budget; this table does not establish equal-test- benefit superiority. Nor does zero penultimate drift under head-only editing guarantee unchanged predictions. These controls make explicit both the restricted scope of the local optimum and the limits of representation RMS as a behavioral preservation metric.

B.10 Predicting the value of the drift metric

Equation (10) concerns squared regularized cost, whereas the empirical endpoint is unregularized RMS. Set $u = H_\lambda^{-1} b$ and normalize both directions to unit linear target gain: $\widetilde{w}_b = b/(b^\mathsf{T} b)$ and $\widetilde{w}_\star = u/(b^\mathsf{T} u)$. The corresponding RMS ratio is

$$R_{\mathrm{RMS}} = \sqrt{\frac{\widetilde{w}_b^\mathsf{T} H \widetilde{w}_b}{\widetilde{w}_\star^\mathsf{T} H \widetilde{w}_\star}}. \tag{25}$$

This uses only the saved validation geometry and no fitted calibration. We compare it with the median of the four frozen test drift ratios within each of the eighteen model–allocation configurations from Appendix B.7. Predicted ratios range from 1.055 to 1.204; observed ratios range from 1.049 to 1.195. Let R be the observed projected-gradient/diagnostic RMS ratio and $\widehat{R} = R_{\mathrm{RMS}}$ its prediction. We report

$$e_{\mathrm{excess}} = \frac{|\widehat{R} - R|}{R - 1}. \tag{26}$$

All eighteen observed ratios exceed one. The denominator measures the comparator's additional drift relative to the diagnostic. The median error is 4.31%, and the maximum is 11.31%. The descriptive configuration-level correlation is 0.991; correlation between the six model medians is 0.927. These nested observations do not provide eighteen independent replications. The comparison in Figure 1B is conditional on the existing split, initialization, and local radius; validation matching does not imply equal test gains.

Comparison with a constant prediction. To test whether the geometry adds information beyond a common drift ratio, we hold out all three allocations of each model together. For each held-out model, a constant predictor is the mean observed ratio over the fifteen configurations of the other five models. Its median excess-normalized error across the eighteen held-out configurations is 22.04%, compared with 4.31% for the unfitted geometry prediction. Geometry has lower mean absolute ratio error in every held-out model and lower absolute error in fourteen of eighteen configurations. This comparison measures predictive variation within the study, with six model folds.

B.11 Measurement tolerance and remaining ambiguity

Exact first-order equality can be replaced by a stated measurement tolerance. Let $r > 0$ bound the update-difference norm and let $\tau \geq 0$ be the measurement-discrepancy tolerance. Write $A = A_{\mathcal{E}}$ and $a = a_{\mathcal{U}}$, using Euclidean norms. The largest first-order target disagreement for a difference h between updates is

$$\Omega(\tau, r) := \max_{\|h\| \leq r,\ \|Ah\| \leq \tau} |a^{\mathsf{T}} h| = \inf_{y \in \mathbb{R}^m} \{\tau \|y\| + r \|a - A^{\mathsf{T}} y\|\}. \tag{27}$$

In particular, $\Omega(0, r) = r\|(I - P_{\mathcal{E}})a\|$. This makes explicit the magnitude of what the measurements leave undetermined, rather than only whether an exactly invisible direction exists. Numerical tolerances must specify the scaling of the measurement coordinates.

To prove the dual expression, symmetry lets us maximize $a^{\mathsf{T}} h$. For every y and feasible h, $a^{\mathsf{T}} h = y^{\mathsf{T}} Ah + (a - A^{\mathsf{T}} y)^{\mathsf{T}} h \leq \tau \|y\| + r\|a - A^{\mathsf{T}} y\|$. For $\tau > 0$, introduce $z = Ah$ and apply convex duality to the two norm balls; $h = z = 0$ is strictly feasible, so the bound is attained in the dual value. For $\tau = 0$, projection onto $\ker A$ gives the displayed value, while least-squares projection of a onto $\mathrm{range}(A^{\mathsf{T}})$ gives the same dual value.

There is also a finite-update bound. Suppose both updates have norm at most r, their nonlinear edit-measurement gap is at most τ, and the local second-derivative operator-norm bounds for $C_{\mathcal{E}}$ and $r_{\mathcal{U}}$ are L_C and L_r. Taylor's theorem gives

$$|r_{\mathcal{U}}(\theta_0 + v_1) - r_{\mathcal{U}}(\theta_0 + v_2)| \leq \Omega(\tau + L_C r^2, 2r) + L_r r^2. \tag{28}$$

The factors account for both updates and their remainders. Applying this bound at an empirical radius requires the stated derivative bounds; the measurement sweep below assesses nonlinear ambiguity directly.

B.12 Increasing the number of defining edit measurements

For each of the eighteen original configurations, we keep all sixty edit images fixed and increase the defining stochastic losses to the first 24 and 48 draws from the same deterministic augmentation sequence. We recompute the row-space basis at each budget. These are additional measurements of the same labels. The original locality comparisons retain their twelve-gradient span. Table 7 reports fractions of gradient norm, not squared energy; each configuration has full row rank at all three budgets. For twelve measurements, the mean squared-norm coverage is 9.26%, computed by squaring each configuration's norm fraction before averaging.

Table 7: Measurement richness across eighteen configurations nested in six models. Residual witnesses use validation target gradients; the last column reports their nonlinear test separation.

Defining losses	Mean captured norm	Mean residual norm	Test-gap range
12	30.1%	95.2%	$[2.63, 3.57] \times 10^{-3}$
24	40.2%	91.4%	$[2.57, 3.45] \times 10^{-3}$
48	50.9%	86.0%	$[2.37, 3.29] \times 10^{-3}$

For each budget, apply the positive and negative normalized target-gradient residual at $1/256$ of the reference-update norm, restoring the checkpoint between interventions. With 48 losses, the defining edit-loss RMS gap is at most 2.0×10^{-5}, while the independent-test target gap remains positive in every configuration. These quantities have different units; their numeric ratio is not an effect-size

comparison. The sweep shows that quadrupling the augmentation measurements increases coverage but does not remove target-relevant ambiguity in this setting. It does not establish invisibility to arbitrary per-example outputs or to the expected loss over all augmentations.

B.13 Independent-test transfer with smaller probe budgets

All six models use allocation $1 + 5$, the original twelve-gradient span, sixty edit images, and the fixed bank of 120 semantic reference images. Reduced target and non-target sets are class-stratified, nested subsets of the original 100 and 900 validation probes. Three deterministic subset draws per model provide within-model variation; the full-budget case is evaluated once. Both directions receive the identical subset. For a reduced budget, neither direction construction nor gain matching uses the remaining target probes. The reference bank remains an additional information requirement.

The prescribed validation margin gain is 5×10^{-5} for every budget. We use the original 10^{-3} ridge fraction, scan radii $1/1024$ through $1/64$ of the reference norm, require an increasing initial valid branch with at most 20% target/drift prediction error, and refine the step with the original bisection tolerance. Coefficients, steps and probe IDs are frozen before evaluation on all 200 target and 9,000 non-target test images. Every planned pair is feasible; lower test drift and greater realized gain/RMS favor the metric in all 78 pairs. Table 8 shows model-level summaries rather than treating subset draws as independent replications.

Table 8: Probe-budget transfer. Counts exclude the fixed 120 reference images and sixty edits. Efficiency is test gain/RMS relative to the projected target gradient: the median and range of six within-model medians.

Target probes	Non-target probes	Pairs	Median efficiency	Model range
10	90	18	1.200	$[1.127, 1.322]$
20	90	18	1.181	$[1.079, 1.275]$
20	900	18	1.196	$[1.080, 1.280]$
100	90	18	1.175	$[1.085, 1.242]$
100	900	6	1.178	$[1.086, 1.243]$

Validation gain matching does not equalize test gains. These local updates assess transfer under the stated probe budgets; classification-scale repair and probe requirements across tasks need separate evaluation. A complementary three-model estimation study used five nested draws: ten target probes per class gave mean direction cosine above 0.9 and relative minimum-cost error below 10%, while one non-target probe per class met the corresponding metric-estimation criteria. Target-gradient estimation thus required more probes under these criteria; the six-model study tests nonlinear transfer under explicit data budgets.

B.14 The cost and limits of edit-loss reliability

Let $e = Q^{\mathsf{T}} \nabla \overline{C}_{\mathcal{E}}(\theta_0)$, where $\overline{C}_{\mathcal{E}}$ averages the twelve defining edit losses. For each original configuration, solve

$$\min_{w} w^{\mathsf{T}} M w \quad \text{subject to} \quad b^{\mathsf{T}} w = \rho, \qquad e^{\mathsf{T}} w \le -\kappa\rho \frac{\|e\|}{\|b\|}, \qquad M \in \{H_\lambda, I\}. \tag{29}$$

Here M selects the regularized drift metric or the Euclidean metric, and κ controls the required edit-loss descent. The fixed choices $\kappa = 0$ and 0.05 impose non-increase and a small strict decrease in the linearized mean edit loss, respectively. These comparisons fix the linear target gain to price the edit-loss constraint at the same benefit. If the unconstrained optimum satisfies the edit constraint, retain it. Otherwise, with $B = [b, e]$ and $t = (\rho, -\kappa\rho\|e\|/\|b\|)^{\mathsf{T}}$, the solution is $M^{-1}B(B^{\mathsf{T}}M^{-1}B)^{-1}t$ whenever the constraints are independent. We check the closed form against an independent constrained numerical solver. All configurations are feasible. This constrains the mean defining loss; it does not constrain each example's loss or its predicted class.

For the nonlinear comparison, each direction uses the original frozen 50%-level requested validation gain, the same radius/error rule, and validation-only step matching. All 72 updates are feasible and are evaluated on the same test images outside construction. The full metric has lower test drift and

greater gain/RMS than the Euclidean objective under identical constraints in all eighteen pairs for each κ. With $\kappa = 0.05$, the six model-median efficiency ratios range from 1.066 to 1.130. Relative to the original unconstrained metric optimum at equal linear gain, the regularized cost inflation has median 1.001 and maximum 1.321.

Table 9: Finite-update reliability checks. Each row covers eighteen configurations within six models; loss changes use original units. First-order non-increase alone permits finite nonlinear increases.

κ	Objective	Mean edit-loss change	Loss increases	Accuracy decreases
0	Euclidean	8.94×10^{-7}	10/18	2/18
0	Full metric	-9.41×10^{-5}	6/18	3/18
0.05	Euclidean	-8.14×10^{-5}	0/18	2/18
0.05	Full metric	-1.39×10^{-4}	0/18	3/18

With the strict descent margin, mean edit loss decreases in every update for both objectives. Nevertheless, edit accuracy decreases in three metric updates and two Euclidean updates. The geometric advantage therefore persists under mean-loss reliability, while decision reliability remains a separate empirical requirement.

Evaluation scope. Directions and step sizes are selected using validation data alone. The 54 measurement-budget configurations, 78 probe-budget pairs, and 36 reliability pairs remain nested within the same six trained models. All 228 selected updates were evaluated, with no exclusions. Numerical checks cover checkpoint restoration, basis orthogonality, per-image reductions, and the constrained solutions.

B.15 Independence and uncertainty

The six stochastic CIFAR training runs are the units of replication for the theory suite. Each model shares a checkpoint and nested edit pools across its three allocations. The eighteen identifiability configurations, 72 test-split locality comparisons, and 24 primary retention model–radius pairs are nested within these models. Results and ranges across models summarize the repeated runs. Allocations, gain levels, continuation streams, and test images describe variation within each model.

Sensitivity checks on models 1–3 comprise 36 equal-improvement comparisons and 4,608 random directions. Figure 3A evaluates 90 classes and 900 examples per model in this subset, giving 270 class–run and 2,700 example–run comparisons. Model IDs identify the models consistently across the distribution rows. The plotting random seed 2027 controls only response sampling and jitter (Appendix B.3).

Numerical verification. We compute orthonormal bases for the edit-gradient row spaces using numerically stable orthogonalization. We evaluate all checkpoints in inference mode. We check gradient recomputation, derivatives, basis orthogonality, and projection residuals for the analyzed runs.

B.16 ImageNet-20 selection criteria

The ImageNet-20 natural-adaptation setting does not satisfy the 10-percentage-point forgetting criterion: the largest ResNet target-class drop is 8.6 percentage points, and ViT does not forget. This setting is excluded from natural-forgetting claims. In the controlled setting, classifier corruption lowers target-class accuracy by 23.5 percentage points for ResNet and 16.2 percentage points for ViT without lowering non-target accuracy, isolating the repair mechanism.