REPAIR OR RELOCATION? WHEN FIXING A COMPOSITION BREAKS ITS PARTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Correcting a model's prediction for a combined change can break its previously correct predictions for the constituent changes. Endpoint accuracy alone cannot distinguish these outcomes. We evaluate repair on controlled geometric quartets by tracking constituent preservation on fixed baseline failures, alongside joint successes gained and lost across the full population. On a fresh bank of 2,348 quartets from 873 parent scenes, 76.6–83.4% of raw-coordinate endpoint corrections introduced constituent errors across three seeds. Joint correctness nevertheless improved by 2.9–6.0 percentage points, showing that incomplete repair can coexist with positive net improvement. Geometric inputs increased constituent-preserving repair by 27.4–33.8 points relative to raw inputs. Focal Distillation reduced regression but also suppressed useful corrections under the tested recipe. Rendered-image experiments showed positive joint gains whose magnitude depended on the training recipe. Repair should therefore be assessed through both constituent preservation and the net change in joint correctness.

1 INTRODUCTION

A receiver and a defender can each move without closing a passing lane, yet block it when both move together. Controlled movements on a recorded tracking frame illustrate this interaction (Figure 1a,b; (Bassek et al., 2025)). A model may judge each movement correctly and still miss their combined effect. Correcting the combined prediction then raises a second question: do those earlier correct judgments survive the update?

An endpoint score cannot answer this question because it credits the correction even when related predictions become wrong. The same correction can therefore reflect two different outcomes: preserving the constituent successes or introducing errors in them (Figure 1c,d). Evaluating repair requires following these judgments together, on the same instances before and after an update.

This distinction connects compositionality gaps with prediction regression after model updates. Compositionality gaps identify failures despite successful component predictions (Press et al., 2023), while positive-congruent training addresses newly introduced errors on previously correct examples (Yan et al., 2021). We examine their intersection by tracking the unchanged-label constituents of each composition being corrected. This pairing makes constituent preservation part of the repair assessment.

We study this question with a controlled segment-intersection task that provides exact labels for two separate changes and their combination. Each evaluation instance is a quartet consisting of a parent scene, its two single edits, and their combined edit. We compare updates on a fixed set of baseline failures and measure which corrections preserve both constituent predictions. We also count joint successes gained and lost across the full population, so local repair and overall improvement can be assessed together.

The resulting evaluation distinguishes repair profiles across representations and supervision strategies. We compare raw coordinates with geometric features, examine a regression-control objective, and extend the evaluation to rendered images. A fresh coordinate bank tests whether the observed patterns persist on newly generated parent scenes. The aim is to measure which related judgments survive an endpoint correction and how that repair contributes to overall accuracy.

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

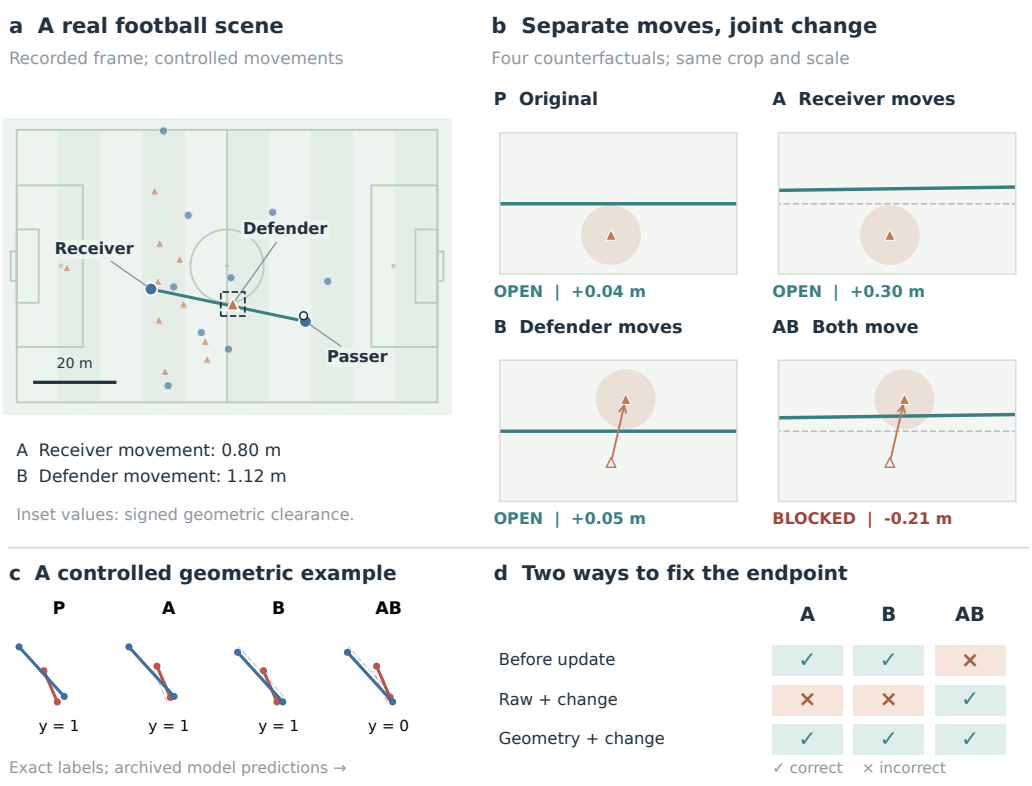

Figure 1: **Endpoint correction and constituent preservation.** (a,b) A receiver and defender can each move without blocking a geometric passing lane, yet block it when both move. Movements are controlled counterfactuals anchored to a recorded frame, not observed later frames or pass outcomes. Insets use the same rotation and scale; discs mark the 0.5 m oracle radius. (c,d) Predictions on a geometric quartet illustrate two endpoint corrections: raw-coordinate change training breaks both constituent predictions, whereas geometric change training preserves them. S8 documents the examples and their selection.

2 A CONTROLLED TEST OF COMPOSED CHANGE

2.1 ORACLE-DEFINED QUARTETS

A scene x contains four planar endpoints grouped into two line segments. A strict intersection predicate provides a binary label $y(x)$. Two edits define $x_A = x + e_A$, $x_B = x + e_B$, and $x_{AB} = x + e_A + e_B$. We retain

$$\mathcal{E} = \{(x, e_A, e_B) : y(x_A) = y(x_B) = y(x), \quad y(x_{AB}) \neq y(x)\}.$$

Eligibility depends only on the oracle, and each state is classified separately. Composition therefore refers to the relationship among inputs and labels, without assuming an internal edit operator. The setting probes excessive invariance to meaningful changes (Jacobsen et al., 2019). Correct constituent predictions, however, do not by themselves establish a learned compositional mechanism.

The coordinate experiment crosses two input representations with clean or change-augmented training. Models receive either eight raw coordinates or ten geometric features: six indexed pairwise distances and four centroid radii. These features encode task structure without directly including the label. All arms use a 64–64–32 MLP trunk and binary readout. Change training adds oracle-labeled single edits and thus requires additional label queries. Supplement S1 provides the generator and complete training recipe.

Table 1: Separate evaluation populations. Unknown parent counts are not inferred from candidate-pool sizes.

Evaluation	Quartets / parents	Role
Coordinate development	237 / 106	Four arms, three seeds, paired analyses
Fresh coordinate confirmation	2,348 / 873	Frozen models; new parent scenes
Archived coordinate comparison	509 / unknown	Historical representation comparison
Historical fresh holdout	176 conditioned cases / unknown	Composition-miss observation
Small rendered-image bank	28 / 19	Exposure-matched positive control
Larger rendered-image bank	685 / 351	Separate recipe and scope check

Distinct evaluation populations (Table 1). We used the 237-quartet bank for diagnostic development, so additional analyses of that bank are exploratory. The archived 509-quartet bank has parent-disjoint provenance relative to development, although its eligible-parent count was not retained. We generated the fresh bank after fixing the models and analysis choices. Results remain separate across banks and seeds; seeds represent repeated fits rather than independent datasets.

2.2 What changed on the same instance?

Let c_A, c_B, c_{AB} denote correctness indicators. Constituent coverage is $A = \Pr(c_A c_B = 1 \mid \mathcal{E})$; joint correctness is $J_3 = \Pr(c_A c_B c_{AB} = 1 \mid \mathcal{E})$. Conditional composition miss is $\mathrm{CCM} = \Pr(c_{AB} = 0 \mid \mathcal{E}, c_A c_B = 1)$. When coverage is nonzero, $J_3 = A(1 - \mathrm{CCM})$. Since CCM's conditioning population can change across models, it cannot alone measure repair.

For baseline f_0, fix $H = \{q \in \mathcal{E} : c_A(f_0) = c_B(f_0) = 1, c_{AB}(f_0) = 0\}$. These are the baseline failures whose two constituent predictions are correct. For candidate f_1, endpoint repair R_{end} is the fraction of H with a correct AB prediction. Full repair R_{full} is the fraction with A, B, and AB all correct. Relocation M is the fraction with a corrected AB prediction but at least one newly incorrect constituent. Thus,

$$R_{\mathrm{end}} = R_{\mathrm{full}} + M.$$

This partition describes prediction outcomes without attributing their underlying causes. Single-atom regression, $110 \to 011/101$, replaces the endpoint error with one constituent error; double-atom regression, $110 \to 001$, increases the error count. Full repair refers only to these three judgments and does not establish general compositional competence. The parent P defines the quartet but is excluded from J_3; requiring its prediction to be correct additionally defines J_4.

The evaluation also covers successes outside H. On the complete bank,

$$\Delta J_3 = \Pr(\text{non-111} \to 111) - \Pr(111 \to \text{non-111}).$$

Both probabilities use the same full-bank denominator. We report new and lost joint successes, rather than measuring only the failures an intervention corrects.

Why marginal scores are insufficient. Consider four baseline-110 quartets. Two hypothetical updates yield U: `011, 101, 110, 110` and V: `111, 111, 010, 100`. Both have A and B accuracies of 3/4, AB accuracy of 1/2, and two newly incorrect atomic judgments. Yet full repair is 0 for U and 1/2 for V. This constructed example proves non-identifiability from those summaries; it is not an experimental model comparison.

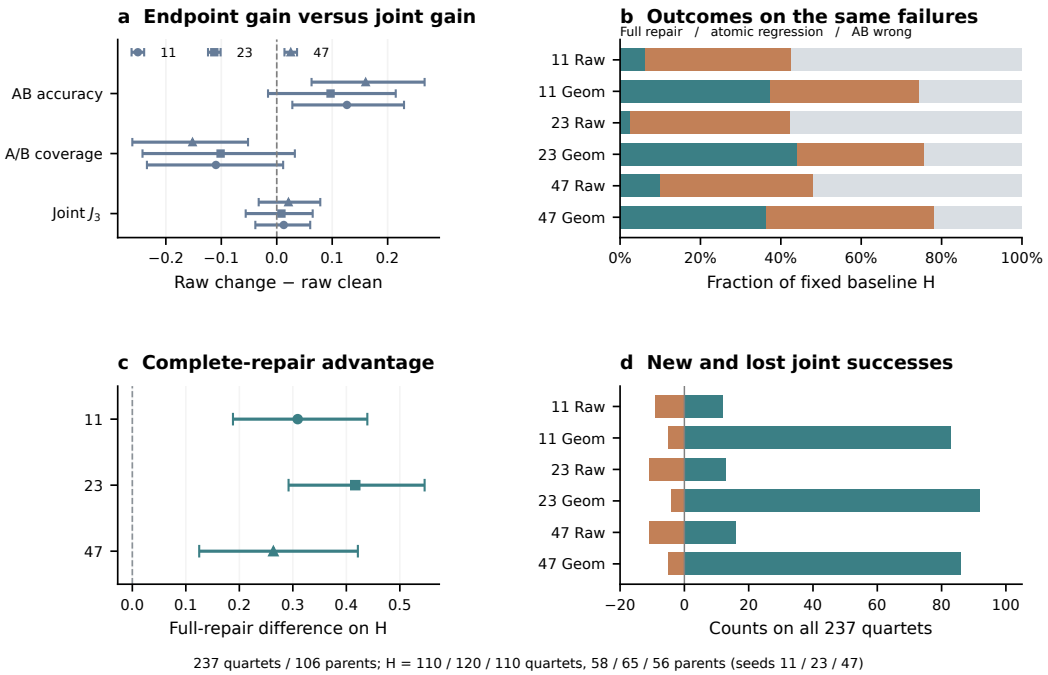

Figure 2: **Endpoint gains coexist with regression.** (a) Paired metric changes for raw change supervision. (b) Raw and geometric change interventions share the same raw-clean H within each seed. Teal denotes full repair, orange constituent regression, and gray an incorrect endpoint. (c) Geometry-minus-raw full-repair differences on H. (d) New joint successes extend right and lost prior successes extend left, using all 237 quartets. Intervals are 95% parent-cluster intervals; seed markers denote training repeats.

Point estimates weight quartets equally. Unless otherwise stated, 95% intervals use 10,000 paired parent-cluster bootstrap draws. Each sampled parent contributes all its quartets. Per-seed intervals remain separate, and empty conditional denominators are undefined. Supplement S3 reports equal-parent sensitivity separately. Matching analyses use their explicitly restricted populations.

3 Endpoint gains versus constituent-preserving repair

The historical fresh holdout recorded 151 misses among 176 compositions with both constituents correct (85.8%). This conditional aggregate motivates the study, but supplies neither an independent-parent count nor a parent-level interval.

The development comparison revealed a divergence between endpoint accuracy and constituent preservation. Raw-coordinate change supervision increased AB accuracy from 0.477–0.519 to 0.574–0.679. Constituent coverage fell from 0.519–0.557 to 0.367–0.456, while J_3 changed from 0.051–0.055 to 0.059–0.076. All three paired ΔJ_3 intervals included zero.

On the fixed H sets, raw change supervision corrected 47, 51, and 53 endpoints out of 110, 120, and 110 baseline failures. Only 7, 3, and 11 became jointly correct. Thus, constituent regression accompanied 85.1%, 94.1%, and 79.2% of **corrected endpoints**; these percentages use corrected endpoints as the denominator. Double-atom regression accounted for 26, 40, and 39 cases, respectively. Many corrections therefore increased the number of errors within a quartet.

Counting all joint successes showed why the net improvement was small. Raw change supervision created 12, 13, and 16 new joint successes but lost 9, 11, and 11 existing successes. Net gains were only 3, 2, and 5 quartets. Geometric change training created 83, 92, and 86 new joint successes while losing 5, 4, and 5. The small raw-coordinate gain therefore reflects both incomplete repairs and losses beyond H. Supplement S2 gives every cell of the eight-state transition matrices.

3.1 Endpoint and joint accuracy can rank candidates differently

Across all three seeds, AB accuracy ranked raw-change above geometry-clean, whereas J_3 reversed that ordering. Geometry-clean's joint-accuracy advantage was 0.063, 0.110, and 0.135. Geometry-change remained the highest-scoring arm under both criteria, so the disagreement affected relative candidate rankings without changing the best arm.

Under the fixed 0.03 absolute tolerance, raw-clean and geometry-clean had similar AB accuracy for seeds 23 and 47. Their complete-repair rates on the raw-clean H nevertheless differed. No pair simultaneously met this tolerance for AB accuracy, three-state average accuracy, and three-state negative-flip rate. The empirical comparison therefore does not establish marginal equivalence or constitute a fresh model-selection trial. The constructed example identifies information missing from marginal summaries, while the observed ordering shows how that distinction can affect candidate assessment.

4 Representation changes the repair profile

Geometry alone yielded $J_3 = 0.131$–0.211, while geometry plus change supervision reached 0.384–0.422. On the shared H sets, geometric change training fully repaired 41, 53, and 40 cases. Its full-repair advantage over raw change was 0.309, 0.417, and 0.264, with positive paired intervals in every seed. Some endpoint corrections nevertheless introduced constituent errors (Figure 2b).

The four-arm contrast

$$I = J_{\text{geom,change}} - J_{\text{geom,clean}} - J_{\text{raw,change}} + J_{\text{raw,clean}}$$

was 0.241, 0.245, and 0.165. All paired parent-cluster intervals excluded zero, supporting representation–supervision complementarity under this training recipe. Hand-designed geometric features provide useful prior structure, and an analytic intersection parser can solve the underlying predicate exactly. The comparison therefore measures how the learner uses available information; it does not show that the task requires a neural solution.

4.1 Incomplete repair persists alongside positive net improvement

After fixing the models and analysis protocol, we generated 4,096 new parent scenes. The same edit construction yielded 2,348 eligible quartets from 873 parents. No duplicates were found against the training and development parents under the checked symmetry-aware coordinate keys. Disjointness from every historical sample remains unverified because sealed banks were not reopened (S9).

Across seeds, 76.6–83.4% of raw-coordinate endpoint repairs involved constituent regression; every 95% parent-cluster interval exceeded 50% (Figure 3c). Geometry-change exceeded raw-change by 28.4–35.0 percentage points in J_3 and by 27.4–33.8 points in full repair on the same H. Thus, both the incomplete-repair pattern and the representation comparison extended to newly generated parents.

Overall improvement was clearer on the fresh bank than on the development bank. Raw change supervision increased J_3 by 2.9–6.0 points, with all three intervals above zero (Figure 3d). New joint successes outnumbered losses by 142, 94, and 67 quartets. Constituent regression thus coexisted with positive net improvement, whereas gains and losses had nearly cancelled on the development bank. The archived 509-quartet comparison independently preserved the raw-clean versus geometry-change ordering, but supplies only two-arm summaries and no eligible-parent count (S9).

4.2 Can a regression-control objective resolve the trade-off?

We evaluated Focal Distillation with logit matching (Yan et al., 2021) using frozen clean teachers and a paired, newly trained baseline. Twelve runs crossed three seeds with weights 0, 0.01, 0.1, and 1. The fixed primary weight, 1, reduced three-state negative flips by 13.0–14.3 points. However, composed-endpoint accuracy fell by 11.2–15.4 points and full repair on H fell by 7.2–8.5 points. The objective preserved more previously correct predictions but also suppressed useful corrections. Both sensitivity settings are reported without post-test selection (S10). This trade-off applies to the tested implementation and task; it does not establish a general limitation of Focal Distillation.

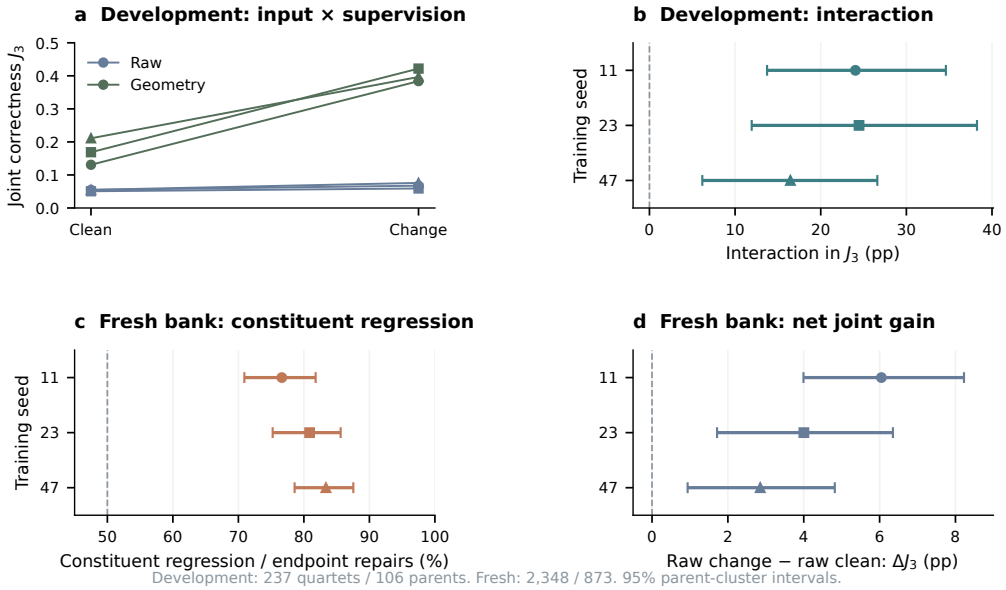

Figure 3: **Incomplete repair can coexist with positive net improvement.** (a,b) The development comparison and representation–supervision interaction use 237 quartets from 106 parents. (c,d) A fresh bank contains 2,348 quartets from 873 parents. Raw change supervision often breaks constituents when fixing the endpoint, yet increases joint correctness overall. Markers distinguish training seeds; intervals are 95% parent-cluster intervals. Banks and seeds are not pooled.

4.3 Alternative explanations

Constants and class direction. Constant predictors achieved $J_3 = 0$ but CCM=1 whenever their atomic-correct subset was nonempty. Their coverage was 0.241 and 0.759, demonstrating why high conditional miss alone is insufficient. Raw-clean accuracy on 512 ordinary source scenes was 0.830–0.867, compared with 0.5 for constants. The eligible bank was imbalanced: 57 quartets required 0→1 and 180 required 1→0. Geometry-change improved J_3 over raw-clean in both strata for every seed, although gains differed by direction.

Matched endpoint difficulty. We matched composed endpoints to oracle-changing single edits from the same parents and label directions. Matching used edit norm, oracle margin, and frozen-baseline confidence, with outcome-blind calipers. The three matched samples contained 163, 160, and 165 pairs from 81, 81, and 82 parents. Every arm's composed-minus-single accuracy interval included zero. This analysis did not establish excess composition-specific endpoint difficulty; it also did not establish equivalence. Residual imbalance and unmatched quartets limit its scope (S4).

Same-instance association. A separate matched-parent analysis retained one H quartet per parent and exchanged updated A/B correctness together. On 25 parent pairs per seed, raw-change endpoint repair co-occurred with atomic regression 16 percentage points, 11 percentage points, and 5 percentage points more often than under paired exchange. The nominal permutation probabilities were 0.0002, 0.0010, and 0.0601. The association varied across seeds, and matching cannot ensure exchangeability with respect to unobserved factors. These exploratory results therefore do not establish a distinct error-transfer mechanism. Threshold selection on a parent split of the exposed development bank also produced mixed changes (S4).

5 Beyond coordinates: useful repair and its scope

To examine whether constituent-preserving repair also occurs with visual inputs, we evaluated rendered RGB images of the geometric task. An ImageNet-initialized ResNet18 (He et al., 2016;

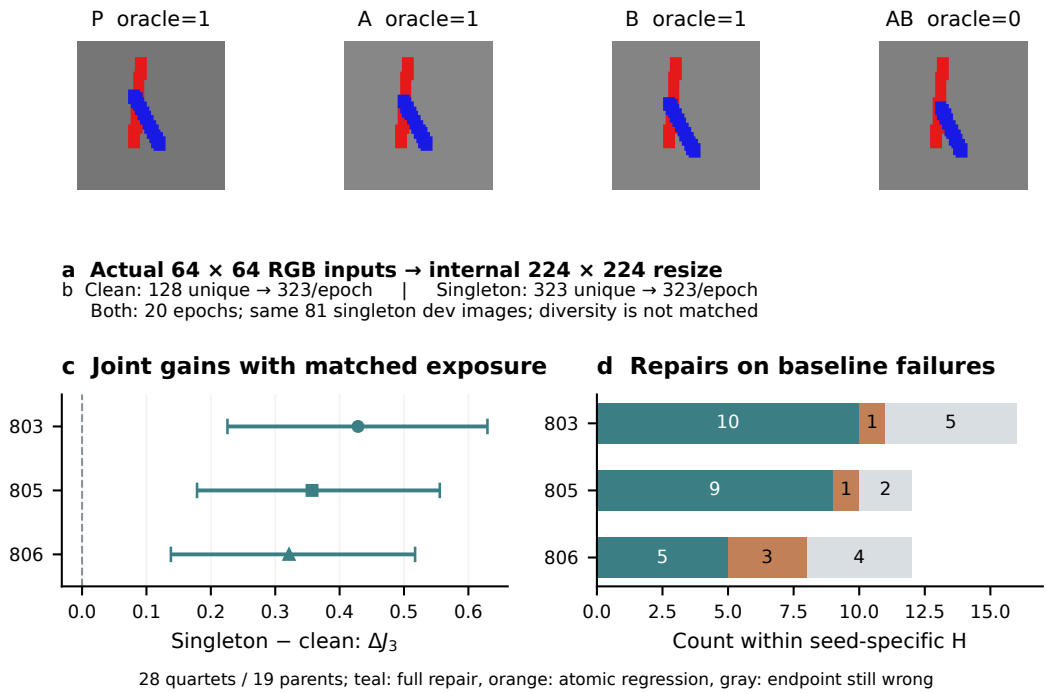

Figure 4: **The protocol also credits useful visual supervision.** Actual RGB inputs are shown without redrawing or contrast changes. The exposure diagram distinguishes unique examples from presentations. Paired ΔJ_3 intervals use parent-cluster resampling. The right panel partitions each clean-baseline H into full repair, atomic regression, and endpoint-still-wrong counts. All results use 28 quartets from 19 parents, with separate training seeds.

Deng et al., 2009) received native 64×64 images resized internally to 224×224. Training used singleton states only, with no composed AB training labels.

The clean-only arm repeated 128 distinct images to match the full-singleton arm's 323 presentations per epoch. Both ran for 20 epochs and selected by BCE on the same 81 singleton development images. Initialization was paired by seed. The comparison matches exposure and model selection while varying unique-example diversity and label coverage.

Full-singleton training increased J_3 from 0.143, 0.250, and 0.214 to 0.571, 0.607, and 0.536. Paired gains were 0.429, 0.357, and 0.321, with positive intervals for all three seeds. On H, full repairs versus relocations numbered 10 versus 1, 9 versus 1, and 5 versus 3. The paired evaluation thus identified substantial constituent-preserving gains from additional visual supervision.

The gains were smaller on a separate rendered bank of 685 distinct quartets from 351 parents, after removing seven duplicate rows. There, exposure-matched direct-classifier gains were 4.5 pp, 2.0 pp, and 2.5 pp, and the seed-805 interval included zero. The large gain on the small bank therefore did not generalize uniformly across the tested recipes. On that larger bank, tuning a geometry-prediction frontend improved J_3 by 18.8–27.3 pp, but remained below direct classification. Different supervision targets and bundled BatchNorm adaptation prevent isolated mechanism attribution (S7). Neither visual bank validates unrestricted natural-image composition.

6 Related Work, Limitations, and Conclusion

Relation to prior work. Systematic recombination is central to compositional generalization benchmarks (Lake & Baroni, 2018; Keysers et al., 2020) and controlled visual reasoning (Johnson et al., 2017). Press et al.'s compositionality gap isolates single-model failures despite successful components. We follow the same component predictions across an update to distinguish correction from constituent-preserving repair. Positive-congruent training addresses general negative flips; our

preservation set is coupled to each corrected composition. MQuAKE instead asks whether edited facts propagate to entailed multi-hop answers (Zhong et al., 2023). In our setting, the constituent labels remain unchanged, so their previously correct predictions provide a natural preservation criterion. The contribution is this instance-level pairing of correction and preservation within established approaches to conditional evaluation and regression measurement.

Limitations. Semantic supervision can expose distinctions missed by ordinary training (Kaushik et al., 2020), but our data do not identify a universal repair mechanism. The development bank was repeatedly examined, the small visual bank contains only 19 parent clusters, and all tasks concern controlled geometry. The matched controls leave alternative explanations unresolved. The tested Focal Distillation settings did not improve repair and preservation together. Cross-renderer and cross-semantic-task generalization remain unestablished. Supplement S5–S7 retains relevant budget, structural, and visual negative evidence. The reproduction supplement provides prediction tables, coordinate checkpoints, scene arrays, and CPU analysis code. Full visual training weights are not bundled, and a permanent external archive is not yet available.

Conclusion. An endpoint correction leaves open whether the model has preserved its related successes. Following the same instances distinguishes full repair from constituent regression, while full-bank accounting measures joint successes gained and lost. These measures separate the repair profiles of the tested interventions and can change their relative rankings. The fresh-bank results show why both views matter: frequent constituent regression can coexist with positive net improvement. Evaluating corrections together with their constituents gives a more complete account of what an update achieves.

AI USE STATEMENT

Generative AI tools assisted the formulation and refinement of the evaluation framework and hypotheses, experimental design, implementation of synthetic-data generators and analysis methods, result interpretation, figure preparation, literature synthesis, and manuscript drafting and editing. Figures were rendered from recorded coordinates and saved numerical predictions, without generative image synthesis. AI-assisted analysis code was checked using constructed examples, independent reconstruction of saved prediction tables, and checkpoint replay. The authors take responsibility for the final text, claims, code, and other artifacts.

ETHICS STATEMENT

The trained-model experiments use synthetic geometric scenes and rendered images. The motivating football illustration uses an existing public professional-soccer tracking dataset, cited to its creators, and controlled geometric counterfactuals. We collected no new human-participant data and make no claims about individual players or actual pass success. The visual model uses publicly available ImageNet-pretrained weights. The controlled task is intended for evaluation of learned behavior; its results should not be treated as a safety certification for deployed models.

REPRODUCIBILITY STATEMENT

Supplement S1 specifies the oracle, generators, preprocessing, data splits, and training recipes. Supplements S2–S4 report complete transitions, weighting, intervals, and alternative-explanation controls; S5–S7 state the scope of the structural and visual evidence. S8 documents the football illustration; S9–S10 specify fresh confirmation and distillation comparisons. The accompanying code supplement provides prediction tables, coordinate checkpoints and scene arrays, generator snapshots, figure code, environment records, and commands for independent metric reconstruction and CPU replay (see its README). Large visual-model checkpoints and pretrained initialization are not bundled, so full visual training reproduction is not claimed.

REFERENCES

Manuel Bassek, Robert Rein, Hendrik Weber, and Daniel Memmert. An integrated dataset of spatiotemporal and event data in elite soccer. *Scientific Data*, 12:195, 2025. doi: 10.1038/

s41597-025-04505-y.

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

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

Jörn-Henrik Jacobsen, Jens Behrmann, Richard S. Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. In *International Conference on Learning Representations (ICLR)*, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2020.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations (ICLR)*, 2020.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *PMLR*, pp. 2873–2882, 2018.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14299–14308, 2021.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, 2023. URL https://aclanthology.org/2023.emnlp-main.971/.

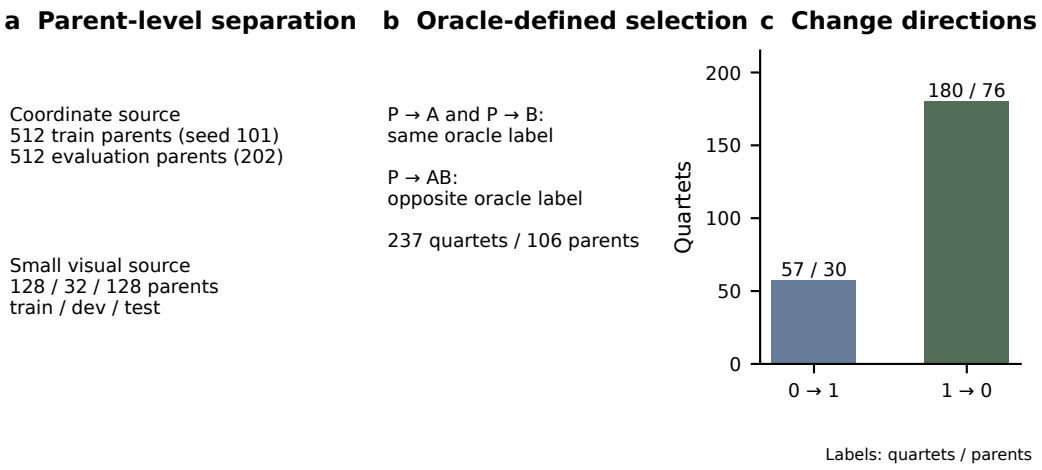

Figure S1: **Generation and evaluation units.** Parent-level splits are distinct from oracle-selected quartet counts. The eligible coordinate bank is directionally imbalanced, despite balanced source parents. No bank is pooled with another.

A Supplementary methods and results

A.1 S1. Generator, populations, and training

Coordinate oracle. For segments (a, b) and (c, d), compute four signed orientation determinants. The label is one only when both orientation pairs have strictly opposite signs. Margin is the minimum absolute determinant. Values at or below 10^{-12} are ambiguous, including collinear, touching, and zero-length degeneracies; invalid shapes and nonfinite coordinates raise errors. The predicate counts proper crossings only and excludes boundary contact.

Parent sampling and quartet construction. Training and ordinary evaluation contain 512 parents each, generated with seeds 101 and 202. Coordinates are independent uniform draws in $[-0.8, 0.8]^2$, accepted until each label has 256 parents with margin at least 0.02. Accepted scenes are shuffled. The development miner uses seed 777 and visits all 512 evaluation parents in a seeded permutation. At each parent it draws four single-endpoint translations of radius 0.10, two whole-segment translations of that radius, and two segment rotations by ±0.15 radians. The archived implementation rotates about the coordinate origin, not the segment centroid. It combines unordered pairs whose individual labels equal the parent and whose joint label differs. All variants retain their parent identity. The bank stores each quartet twice as directed A→AB and B→AB paths; these are aligned by explicit state type and collapsed to one quartet.

The coordinate quartet miner does not apply the single-edit endpoint margin floor to composed states. All 948 quartet states satisfied the strict oracle, with a minimum margin of 0.000168 and no ambiguous labels. No near-boundary cases were removed, leaving the original 237/106 quartet/parent counts. The two direction strata contain 57/30 (0→1) and 180/76 (1→0) quartets/parents. The generator and selection rules are included in the supplement; historical rejection counts are unavailable.

Training recipe. Both input conditions use linear(input,64)–ReLU–linear(64,64)–ReLU–linear(64,32)–linear(32,1). There is no activation between the last two linear maps. Raw and geometric models contain 6,849 and 6,977 parameters. Geometry concatenates indexed distances (01,02,03,12,13,23) and four distances to the centroid; the legacy name `rel12` still denotes ten inputs. Features are standardized using clean training means and standard deviations plus 10^{-8}. Change states use those same statistics. Initialization uses PyTorch seeds 11,23,47 before constructing each model. Raw and geometric input dimensions differ, so this is not identical tensor initialization across representations.

All four conditions use Adam, learning rate 0.01, default betas (0.9,0.999), epsilon 10^{-8}, zero weight decay, and 300 full-batch optimization steps. Clean loss is mean binary cross-entropy on 512 scenes.

Change loss adds, with weight one, mean BCE on 1,534 mined single-edit states. Each pool's mean is computed separately; this is not mean BCE on the concatenated 2,046 rows. Mining uses seed 5, a maximum of 12 valid flips per parent, and endpoint margin at least 0.03. The deterministic candidate grids and random candidates are included in the code snapshot. Every retained flip is oracle-labeled. All arms save the final step, without development early stopping. The arms match optimizer steps and the clean component, but differ in state presentations and FLOPs. They therefore do not provide a comparison of compute efficiency.

Table S1: Coordinate training populations and optimization budgets.

Condition	Unique clean	Additional labeled edits	Steps	Presentations	Parameters
Raw clean	512	0	300	153,600	6,849
Raw change	512	1,534	300	613,800	6,849
Geometry clean	512	0	300	153,600	6,977
Geometry change	512	1,534	300	613,800	6,977

Visual generation and training. The small image bank has disjoint train/dev/test parent sets of 128/32/128, using parent seeds 832701/832702/832703. Observation seeds are 832711/832712/832713; quartet mining and observation seeds are 832721/832722. Each training parent contributes a clean state and up to two label-changing single edits. Sixteen fixed trials cycle through radii 0.10,0.20,0.35,0.50; candidates require margin at least 0.02 and 20 changed pixels. Native64 rendering draws red and blue segments on per-image uniform gray backgrounds; endpoints within each segment are canonically ordered before rendering. The quartet miner requires margins at least 0.005, with at most two accepted quartets per test parent. It produces 28 quartets, 112 state images, and 19 eligible parents. Train/dev contain no AB labels. Backgrounds vary across quartet states, so this is not a background-controlled causal intervention.

The resulting train/dev/test singleton counts are 323/81/319. Both compared direct models use pretrained ResNet18, global pooling and a linear binary head, Adam at 3×10^{-4}, batch size 32, 20 epochs, and minimum dev BCE selection on the same 81 singleton images. All backbone parameters are trainable. Inputs are bilinearly resized to 224×224 with `align_corners=False` and ImageNet normalization, identically at training and evaluation. Seeds are 803/805/806. Clean exposure is the deterministic repeated clean-index vector of length 323, shuffled each epoch. Both arms receive 6,460 presentations, but unique samples and label coverage differ. Parent and image splits were checked, and paired initial-backbone hashes matched. The larger visual experiment uses a separate recipe, described in S7, and is not merged with this bank.

Population provenance. The 237 development bank, 509 archived confirmation, 151/176 historical conditional holdout, and both visual banks have different roles. The 509 summary records 509 quartets but not the number of eligible parents; 1,000 requested parent candidates is not an eligible-parent count. The historical 176 denominator is conditional on constituent correctness and has no recoverable cluster count in the cited summary. Neither raw sealed bank was reread for this revision.

A.2 S2. Complete correctness transitions and net change

All 64 transitions are provided for each arm and seed in the machine-readable table. The figure displays the two principal interventions relative to the same-seed raw-clean baseline. Row/column states denote correctness in A,B,AB order. Counts, including zeros, cover the entire 237-quartet bank.

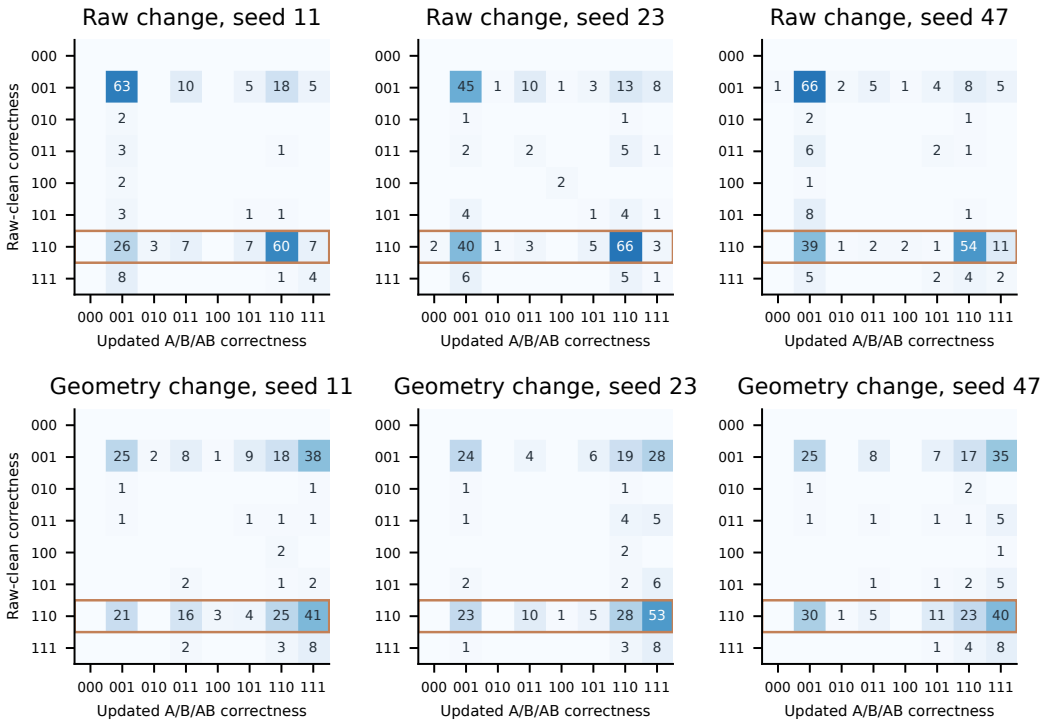

Figure S2: **Complete eight-state transitions.** Orange outlines identify the baseline-110 row. The baseline-111 row separately exposes losses of previous joint success. Each panel contains 237 quartets from 106 parents; seed repetitions are not pooled.

Table S2: Fixed-baseline repair counts by candidate and seed.

Seed	Candidate	H	Full	One atom	Two atoms	AB wrong
11	Raw change	110	7	14	26	63
11	Geometry clean	110	15	13	15	67
11	Geometry change	110	41	20	21	28
23	Raw change	120	3	8	40	69
23	Geometry clean	120	28	12	8	72
23	Geometry change	120	53	15	23	29
47	Raw change	110	11	3	39	57
47	Geometry clean	110	17	7	15	71
47	Geometry change	110	40	16	30	24

All repair counts above use H, whereas the next table uses the complete bank. The net change equals (new joint successes − lost joint successes)/237. Baseline-111 counts are 13,12,13 for seeds 11,23,47.

Table S3: New and lost joint successes across the full development bank.

Seed	Candidate	New 111	Lost 111	Paired ΔJ3 [95% CI]
11	Raw change	12	9	0.013 [-0.039, 0.060]
11	Geometry clean	31	13	0.076 [0.008, 0.144]
11	Geometry change	83	5	0.329 [0.235, 0.426]
23	Raw change	13	11	0.008 [-0.056, 0.065]
23	Geometry clean	40	12	0.118 [0.038, 0.196]
23	Geometry change	92	4	0.371 [0.268, 0.474]

Seed	Candidate	New 111	Lost 111	Paired ΔJ3 [95% CI]
47	Raw change	16	11	0.021 [-0.033, 0.079]
47	Geometry clean	42	5	0.156 [0.090, 0.224]
47	Geometry change	86	5	0.342 [0.248, 0.438]

A.3 S3. Metric definitions, weighting, and candidate comparisons

Quartet-weighted point estimates and parent-cluster intervals target the ratio of pooled row counts. Each of 10,000 bootstrap draws samples the same number of parent clusters with replacement, retaining all rows in selected clusters. Random seed is 92501. Conditional H intervals sample H parents; CCM and relocation-share bootstrap denominators are recalculated within every draw. All these main-bank draws were valid. The parent-equal sensitivity instead averages within each parent and then over parents. It is a different estimand. Intervals are per-seed descriptive intervals, without a three-seed normal approximation or multiplicity adjustment.

Table S4: Endpoint, coverage, and joint metrics by arm and seed.

Seed	Arm	AB	Coverage	J3	CCM	Parent-equal J3
11	Raw clean	0.519	0.519	0.055	0.894	0.055
11	Raw change	0.646	0.409	0.068	0.835	0.086
11	Geometry clean	0.451	0.675	0.131	0.806	0.165
11	Geometry change	0.764	0.595	0.384	0.355	0.405
23	Raw clean	0.477	0.557	0.051	0.909	0.052
23	Raw change	0.574	0.456	0.059	0.870	0.055
23	Geometry clean	0.468	0.696	0.169	0.758	0.188
23	Geometry change	0.747	0.671	0.422	0.371	0.404
47	Raw clean	0.519	0.519	0.055	0.894	0.064
47	Raw change	0.679	0.367	0.076	0.793	0.087
47	Geometry clean	0.494	0.709	0.211	0.702	0.213
47	Geometry change	0.789	0.603	0.397	0.343	0.412

The table below reports J3 intervals under quartet and equal-parent weighting for every arm. A, B, AB, coverage, state accuracy and CCM intervals are also available in the machine-readable metrics file.

Table S5: Quartet-weighted and parent-equal joint-score intervals.

Seed	Arm	Quartet J3 [95% CI]	Parent-equal J3 [95% CI]
11	Raw clean	0.055 [0.019, 0.102]	0.055 [0.019, 0.099]
11	Raw change	0.068 [0.034, 0.107]	0.086 [0.040, 0.138]
11	Geometry clean	0.131 [0.083, 0.183]	0.165 [0.104, 0.231]
11	Geometry change	0.384 [0.292, 0.478]	0.405 [0.321, 0.492]
23	Raw clean	0.051 [0.013, 0.103]	0.052 [0.017, 0.096]
23	Raw change	0.059 [0.027, 0.096]	0.055 [0.022, 0.093]
23	Geometry clean	0.169 [0.115, 0.228]	0.188 [0.124, 0.256]
23	Geometry change	0.422 [0.321, 0.522]	0.404 [0.317, 0.493]
47	Raw clean	0.055 [0.022, 0.097]	0.064 [0.025, 0.110]
47	Raw change	0.076 [0.036, 0.126]	0.087 [0.042, 0.138]
47	Geometry clean	0.211 [0.142, 0.283]	0.213 [0.146, 0.286]
47	Geometry change	0.397 [0.305, 0.492]	0.412 [0.329, 0.502]

Repair fractions. Full and relocation rates below use H; their sum is endpoint repair. The relocation share has endpoint repairs as its denominator, and must not be read as a proportion of H.

Table S6: Full repair, relocation, and the conditional relocation share.

Seed	Arm	Full [95% CI]	Relocation [95% CI]	Relocation/end
11	Raw change	0.064 [0.018, 0.125]	0.364 [0.238, 0.509]	0.851
11	Geometry change	0.373 [0.262, 0.495]	0.373 [0.212, 0.522]	0.500
23	Raw change	0.025 [0.000, 0.055]	0.400 [0.267, 0.542]	0.941
23	Geometry change	0.442 [0.320, 0.569]	0.317 [0.179, 0.464]	0.418
47	Raw change	0.100 [0.035, 0.183]	0.382 [0.238, 0.539]	0.792
47	Geometry change	0.364 [0.237, 0.510]	0.418 [0.245, 0.574]	0.535

Candidate interpretation. All six within-seed arm pairs were examined under an absolute 0.03 closeness rule. Only raw-clean versus geometry-clean in seeds 23 and 47 met the endpoint criterion. No pair met endpoint, state accuracy and three-state-row negative-flip criteria simultaneously. The tolerance defines numerical closeness and does not test statistical equivalence. Across all seeds, the endpoint and J3 rankings differed between raw-change and geometry-clean; geometry-change remained first. These rankings describe the development bank; they do not show that held-out model selection would choose a different best method.

Table S7: Paired candidate contrasts on the same development quartets.

Seed	Geometry-clean minus raw-change	ΔAB	ΔJ3 [95% CI]
11	Same 237 quartets	-0.194	0.063 [0.004, 0.123]
23	Same 237 quartets	-0.105	0.110 [0.047, 0.179]
47	Same 237 quartets	-0.186	0.135 [0.049, 0.217]

A.4 S4. Targeted alternative-explanation controls

Constants and direction. The training label distribution is balanced, so the prespecified majority tie-break chooses label zero. Both constants score 0.5 on the 512 ordinary evaluation scenes. Constant zero and one have development coverage 57/237 and 180/237, respectively, J3=0, and CCM=1 on their nonempty atomic-correct subsets. A large conditional miss rate alone therefore cannot identify a composition-specific mechanism. The table below reports J3 separately for each change direction.

Table S8: Joint correctness stratified by oracle-change direction.

Seed	Arm	0→1 J3 (57/30)	1→0 J3 (180/76)
11	Raw clean	0.018	0.067
11	Raw change	0.053	0.072
11	Geometry clean	0.035	0.161
11	Geometry change	0.193	0.444
23	Raw clean	0.053	0.050
23	Raw change	0.088	0.050
23	Geometry clean	0.140	0.178
23	Geometry change	0.228	0.483
47	Raw clean	0.053	0.056
47	Raw change	0.105	0.067
47	Geometry clean	0.035	0.267
47	Geometry change	0.246	0.444

Difficulty matching. Of 126,151 stored single edits, 3,295 change the oracle label. Single endpoints were originally retained only at margin $\geq$0.005, unlike all composed endpoints. For each seed, targets are processed in increasing qid. Candidates share the parent and label direction and are used at most once. Covariates are log edit norm, log endpoint margin, and absolute raw-clean endpoint logit. Scaling uses the pooled candidate/target standard deviation. Match calipers are

maximum absolute standardized coordinate distance 1.5 and root-mean-square distance 0.75. The nearest eligible candidate is selected, breaking ties by stored order. Only baseline information enters matching, never candidate outcomes. The matched comparison applies only to common support; the principal analysis retains unmatched quartets.

Table S9: Matching support and residual covariate imbalance.

Seed	Matched/237	Parents	Unmatched	Residual SMD: norm/margin/confidence
11	163/237	81	74	0.203 / -0.090 / 0.056
23	160/237	81	77	0.186 / -0.078 / 0.018
47	165/237	82	72	0.186 / -0.117 / 0.025

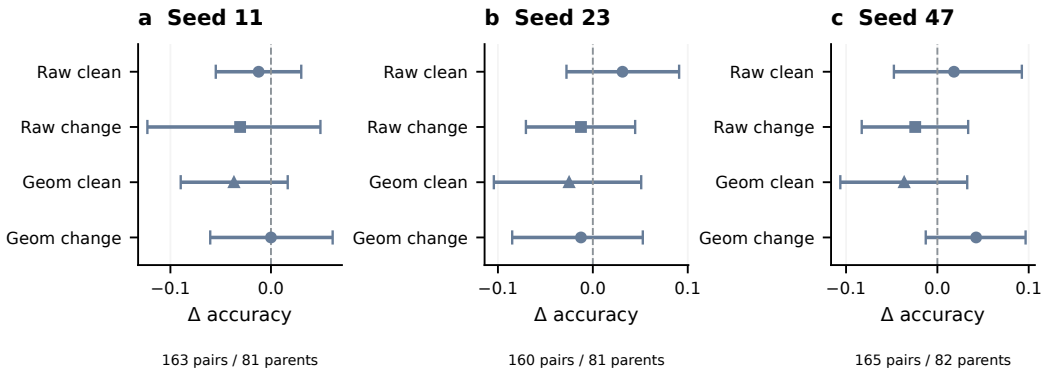

Figure S3: **Matched endpoint accuracy differences.** Point estimates and 95% parent-cluster intervals compare composed endpoints with matched oracle-changing single edits. Different seeds have different matches because matching uses their frozen baseline confidence. Intervals containing zero do not prove equivalence.

Table S10: Composed and matched single-edit endpoint accuracy.

Seed	Arm	Composed accuracy	Single accuracy	Difference [95% CI]
11	Raw clean	0.540	0.552	-0.012 [-0.055, 0.030]
11	Raw change	0.693	0.724	-0.031 [-0.123, 0.049]
11	Geometry clean	0.417	0.454	-0.037 [-0.090, 0.017]
11	Geometry change	0.804	0.804	0.000 [-0.060, 0.061]
23	Raw clean	0.506	0.475	0.031 [-0.028, 0.091]
23	Raw change	0.613	0.625	-0.013 [-0.071, 0.045]
23	Geometry clean	0.438	0.463	-0.025 [-0.104, 0.051]
23	Geometry change	0.794	0.806	-0.013 [-0.085, 0.053]
47	Raw clean	0.521	0.503	0.018 [-0.048, 0.092]
47	Raw change	0.679	0.703	-0.024 [-0.083, 0.034]
47	Geometry clean	0.467	0.503	-0.036 [-0.106, 0.033]
47	Geometry change	0.830	0.788	0.042 [-0.013, 0.097]

Same-instance exchange control. From each H parent retain its smallest qid. Match different parents into disjoint pairs, exact on oracle direction and using log composed norm, log endpoint margin, mean absolute A/B baseline logit, and absolute AB baseline logit. Covariates are standardized within this parent-selected set, with the same 1.5/0.75 calipers. Sort eligible edges by distance and greedily retain disjoint pairs. Each seed yields 25 pairs (50 distinct parents), from 58/65/56 H parents. Missing support is 8/15/6 parents, respectively. Fix the AB repair indicator z; exchange updated A/B correctness as an intact pair. The statistic is mean z times the atomic-regression indicator. Its

difference from the paired-exchange expectation measures selected-subset association. Ten thousand identity-or-swap draws use seed 92501 plus the training seed. Bootstrap intervals resample matched parent pairs as blocks, not the complete H population.

Table S11: Exploratory matched-parent exchange control.

Seed	Arm	Observed	Exchange mean	Excess [95% CI]	Nominal p
11	Raw change	0.440	0.280	0.160 [0.110, 0.210]	0.0002
11	Geometry change	0.320	0.290	0.030 [0.000, 0.060]	0.2534
23	Raw change	0.480	0.370	0.110 [0.060, 0.160]	0.0010
23	Geometry change	0.300	0.270	0.030 [0.000, 0.060]	0.2526
47	Raw change	0.400	0.350	0.050 [0.010, 0.090]	0.0601
47	Geometry change	0.340	0.300	0.040 [0.010, 0.080]	0.1206

The six exploratory probabilities are unadjusted for multiple comparisons. Bootstrap intervals and paired randomization tests use different constructions and need not agree. For example, raw-change seed 47 has a positive percentile interval but p≈0.060. Residual confounding and restricted matching support limit the analysis, and the permutation results are not significant across all seeds. The supplement includes the matched pairs and distances.

Threshold sensitivity. Parent IDs modulo three define a selection subset (remainder zero) and a disjoint descriptive evaluation subset. Both come from the previously exposed development bank. Thresholds are fixed to $-4,-2,-1,0,1,2,4$. Maximize selection J3; break ties by proximity to zero, then the smaller threshold. Compare the selected threshold with the original zero threshold only on the other parents. Thresholds are selected without evaluation-subset outcomes, but both subsets belong to the previously examined development bank.

Table S12: Threshold selection and evaluation on exposed development subsets.

Seed	Arm	Selected threshold	Evaluation ΔJ3 [95% CI]
11	Raw clean	0	0.000 [0.000, 0.000]
11	Raw change	-4	0.027 [-0.025, 0.083]
11	Geometry clean	4	0.054 [-0.028, 0.132]
11	Geometry change	4	-0.135 [-0.257, -0.013]
23	Raw clean	2	-0.007 [-0.072, 0.043]
23	Raw change	-4	0.061 [-0.008, 0.146]
23	Geometry clean	4	0.007 [-0.101, 0.113]
23	Geometry change	2	-0.122 [-0.251, 0.008]
47	Raw clean	-1	0.000 [-0.034, 0.032]
47	Raw change	-2	-0.020 [-0.067, 0.021]
47	Geometry clean	-1	-0.020 [-0.062, 0.015]
47	Geometry change	-1	-0.007 [-0.044, 0.027]

A.5 S5. Budget and supervision controls from separate experiments

The following archived experiments examine supervision and budget effects on separate banks. Their results supplement the controls on the principal 237-quartet comparison. The retained-label study compared six-distance and raw inputs after mining complete flip pools. At training sizes N and 4N, it retained 383/1,534 and 1,597/6,390 labels in 25% arms, following 2,801 and 11,955 mining attempts. Evaluation used 1,229 quartets from 250 parents. Only retained labels were reduced; oracle queries were still required during mining. Its 648 continuation arms share this evaluation bank and are not independent replications.

A capacity/budget comparison on separate T1 and T2 tasks examined one typed width over a three-learning-rate by two-step grid. Six-distance-minus-typed joint-score contrasts remained positive across three seeds, but a two-thread rerun changed exact estimates. The comparison covers only

the architectures and training recipes tested. The separate banks contain 2,724/322 and 3,443/369 quartets/parents. They do not establish a positive interaction on every task.

A prospective readout diagnostic met its primary direction in 7/8 cells, falling short of its registered all-cell criterion. That comparison used previously generated banks and provides no fresh-bank confirmation.

A.6 S6. Structural evidence and limits

A constructed T1 example has opposite oracle labels, margin 0.09375, and identical independently sorted feature vectors in both float32 and float64. Unsorted and legal whole-orbit features distinguish that particular pair. The example establishes an information-loss case for independent sorting, without showing how often such collisions occur. Source searches found positive sampled opposite-label distances (0.0471 in a finite bank and 0.0335 after a bounded local search). The observed minima do not provide positive lower bounds on the global minimum.

Legal whole-orbit representations preserve role correspondence by applying one allowed permutation across the feature vector. Archived clean-training results favor that source representation, but flip-side differences against sorting are $+0.0654, +0.0005, -0.0311$. No stable three-seed flip advantage is established. Historical mixed row-weighted estimates and parent-equal intervals are not used as inferential support. Independent replay reproduced results from all 36 selected checkpoints and 21 paired contrasts. It also corrected two diagnostic fields: an older CCM field stored conditional success, and some repair records mislabeled a raw-change baseline as raw-clean. The corrected joint-score contrasts remain seed-dependent, so these checks do not establish a superior repair mechanism.

The visual interaction-head advantage was also unstable: interaction-minus-direct J3 contrasts were $+0.179, -0.143, 0.000$ in one recipe and $-0.250, -0.036, 0.000$ after an input-pooling correction. The wider-network sampling study simultaneously changed parameters and computation. Its exploratory fraction-of-gap interval crossed 0.5; compute, capacity and sampling contributions were not independently identified. The experiments therefore do not isolate a sampling mechanism or establish that an interaction architecture is generally necessary.

A.7 S7. Visual evidence and local prediction geometry

Small visual bank. Main Figure 4 displays the original RGB arrays using nearest-neighbor enlargement, without contrast adjustment. The illustrative quartet is chosen by fixed seed 803 and increasing parent/identifier, requiring baseline 110 and candidate 111. It is not selected for effect size. The following intervals are recomputed from archived predictions with the same parent-cluster estimator used in the revision.

Table S13: Small visual-bank joint scores and paired gains.

Seed	Clean J3	Singleton J3	ΔJ3 [95% CI]	H / parents
803	0.143	0.571	0.429 [0.226, 0.630]	16 / 12
805	0.250	0.607	0.357 [0.179, 0.556]	12 / 9
806	0.214	0.536	0.321 [0.138, 0.517]	12 / 9

Larger visual bank. The 16-cell experiment originally stored 692 rows from 351 parents. Seven repeated parent/unordered-exact-edit pairs were removed by retaining the first occurrence without model outcomes, giving 685 distinct quartets. All conditions use that same primary denominator. Each trained cell receives 50,640 presentations over 20 epochs. Clean-only repeats 1,024 unique images to 2,532 presentations per epoch; full-singleton uses 2,532 distinct rows. Both select on 333 singleton dev images. Direct classifiers select by BCE; geometry frontends select by geometry MSE. The latter have privileged geometric targets, so equal exposure does not imply equal supervision information.

Table S14: Larger visual-bank joint scores.

Condition	803	805	806
Direct clean	0.864	0.873	0.892
Direct singleton	0.909	0.893	0.917
Frozen geometry frontend	0.182	0.153	0.178
Tuned geometry frontend	0.455	0.342	0.382
Random head, true geometry	0.000	0.006	0.013

The trained frozen head scored 0.7942 on true geometry. This single deterministic reference is shared across seeds. Random-head controls use true geometry; they are not matched controls on noisy learned frontend outputs. An earlier implementation disabled backbone gradients in both modes and changed BatchNorm buffers during the frozen condition; those runs are excluded. In the corrected runs reported here, frozen parameters and buffers remained unchanged, while tuned backbones received nonzero gradients and parameter updates. Tuning still combines weight learning with BatchNorm adaptation.

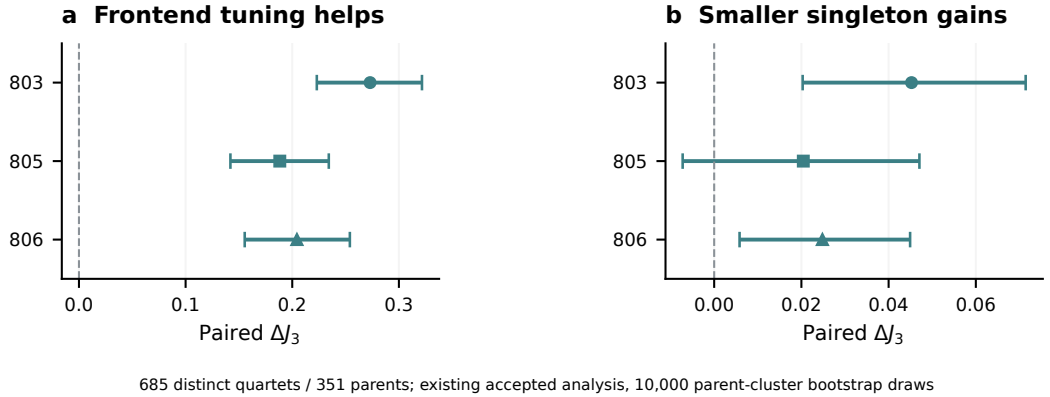

Figure S4: **Larger-bank contrasts.** Tuning gains are positive, but direct singleton-versus-clean gains are smaller than in the small bank, with the seed-805 interval crossing zero. Intervals use 10,000 parent-cluster draws over 685 distinct quartets from 351 parents. No banks or seed-level samples are pooled.

Table S15: Larger visual-bank paired contrasts.

Seed	Contrast	ΔJ3 [95% CI]
803	Tuned − frozen	0.273 [0.223, 0.322]
803	Singleton − clean	0.045 [0.020, 0.071]
803	Direct − tuned frontend	0.454 [0.411, 0.497]
805	Tuned − frozen	0.188 [0.142, 0.234]
805	Singleton − clean	0.020 [-0.007, 0.047]
805	Direct − tuned frontend	0.552 [0.509, 0.595]
806	Tuned − frozen	0.204 [0.155, 0.254]
806	Singleton − clean	0.025 [0.006, 0.045]
806	Direct − tuned frontend	0.534 [0.490, 0.580]

The larger bank is a separate recipe, not a matched replication of the 28-quartet experiment. Cross-bank physical-parent disjointness is not established here. Cross-renderer transfer, BN-only adaptation, geometry-error-matched interventions and cross-predicate reuse remain untested. Frontend gains under this recipe therefore leave the value of modularity and the learnability of the interface unresolved.

Selected edit plane. For the geometric quartet in main Figure 1c,d, we evaluate 101×101 points in $x + \alpha e_A + \beta e_B$, with $\alpha, \beta \in [0, 1]$. Features are recomputed from each scene and use the original checkpoint normalization. Four-corner decisions match the stored quartet predictions. Oracle-ambiguous grid points are masked. Predictions are computed independently at every grid point, without interpolation between corners.

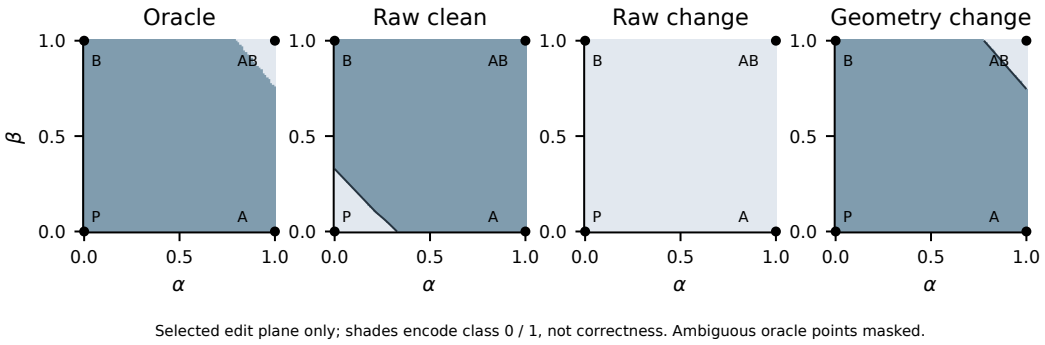

Figure S5: **Local prediction geometry.** The oracle and three models are evaluated on one selected edit plane. Colors indicate predicted classes rather than correctness, and contours indicate zero logits. The figure illustrates boundary arrangements associated with the observed repair outcomes; it does not identify training causality or establish a population-level mechanism.

A.8 S8. Figure 1 examples and provenance

Recorded frame and geometric proxy. Figure 1a redraws numerical tracking coordinates from the dataset of Bassek et al. (2025), data DOI: 10.6084/m9.figshare.28196177. The selected frame is match J03WOY, period firstHalf, timestamp 1662835263560 ms. All players and the ball are drawn at recorded positions; no photograph or generative image was used. The highlighted receiver is a same-team candidate, not the observed recipient of that pass. Player markers encode roles and do not represent physical player size.

The passing-lane oracle is open when no eligible defender is less than 0.5 m from the passer-to-receiver segment. A defender within 1 m of either endpoint is excluded by the original oracle. All eleven opponents contribute to the label, although the local inset centers on the constraining defender. These labels are geometric proxies, not observed or predicted pass success.

Controlled movements. The example is an existing random-search witness anchored to the recorded frame. A moves only the candidate receiver by 0.803 m; B moves only the focal defender by 1.124 m. Both stay within the archived 1.214 m displacement cap, derived from the match's empirical 0.2-second motion envelope. All other player positions are fixed. A, B, and AB are controlled counterfactuals; none is claimed to be the subsequent observed full frame. The cap is an empirical envelope, not proof of joint dynamic feasibility.

For readability, selection used only archived coordinates and oracle labels. We retained open-to-blocked witnesses whose focal defender stayed more than 1.5 m from both endpoints in every state. Its projection also had to remain between 0.1 and 0.9 of the lane length. Among these candidates, we maximized the minimum absolute oracle margin, breaking ties by match, event, and receiver identifier. This avoids an illustration driven by the endpoint-exclusion rule and does not select model outcomes. It is an illustrative case, not a prevalence estimate.

The four insets use the same rigid rotation and magnification around the focal defender. Dashed lines retain the original lane; the circle denotes the 0.5 m oracle radius, and the arrow denotes B. Clearance is the minimum eligible defender-to-segment distance minus 0.5 m. The displayed states were rechecked with a scalar geometric calculation before drawing.

Table S16: Football illustration, with all other players held fixed.

State	Receiver changed	Defender changed	Oracle label	Clearance (m)
P	no	no	1	+0.0403
A	yes	no	1	+0.2988
B	no	yes	1	+0.0546
AB	yes	yes	0	-0.2075

Recorded-movement comparison. The archived audit separately crossed identity-linked receiver and defender movements from later frames within 0.2 s. The passer and the other ten defenders remained at their starting positions, so these are also controlled combinations, not observed full-frame transitions. Eligible combinations occurred in 8/1,449 candidate situations, involving 8/162 audited events and 10/7,117 admissible frame pairs. In contrast, guided/random search found witnesses in 670/1,449 candidate situations. Search constructibility, situation frequency and frame-pair frequency use distinct numerators and must not be interchanged. These counts describe the sampled situations, not all passes in the three matches.

Geometric model example. Figure 1c,d retains the previously selected quartet, parent 16 and identifier 191, with training seed 11. Selection used a fixed parent/identifier ordering among cases with raw-clean correctness 110, raw-change 001 and geometry-change 111. These bits describe correctness on A/B/AB, not class labels; the oracle labels are 1/1/0. The parent constructs the instance but is excluded from the three-judgment repair criterion. Saved logits and scene arrays supply every displayed value. The football panels contain oracle labels only; model predictions come from the separate geometric task.

A.9 S9. Fresh coordinate confirmation

Generation and separation. Models and analysis choices were fixed before generating 4,096 new parent scenes. Parent and edit seeds were 2026092601 and 2026092602. Applying the same oracle-defined edit construction considered 114,688 edit pairs and retained 2,348 quartets from 873 eligible parents. The generator recorded 9,613 pairs with a changed constituent label and 102,727 with an unchanged composed label; no ambiguous pairs remained. Symmetry-aware coordinate keys, quantized at 10^{-6}, found no overlap with the allowed clean training and development parent sets. Overlap with sealed historical banks remains unverified because those banks were not reopened.

The primary confirmation compares the twelve frozen historical four-arm models. The twelve distillation models in S10 were trained before bank generation and evaluated separately. There was no choice of weight, epoch, checkpoint or stopping rule based on fresh-bank outcomes. Training seeds are repeated fits on the same bank, not separate data replications. Intervals resample parent clusters 10,000 times and retain quartet-weighted estimands; no cross-seed pooling is used.

Table S17: Fresh-bank confirmation; differences are percentage points and intervals are 95% parent-cluster intervals.

Seed	H quartets / parents	Regression / endpoint repairs (%)	Geometry − raw ΔJ3	Geometry − raw full repair / H
11	1246 / 513	76.6 [70.9, 81.8]	28.4 [24.8, 32.2]	27.4 [22.3, 32.6]
23	1273 / 515	80.9 [75.2, 85.6]	35.0 [31.6, 38.5]	33.8 [29.0, 38.7]
47	1278 / 517	83.4 [78.6, 87.5]	32.4 [29.0, 35.8]	31.6 [27.0, 36.2]

All contrasts above use the same raw-clean H within each seed. Regression share conditions on corrected AB, whereas full-repair rate conditions on all H. The raw-change net benefit below uses the complete fresh bank.

Table S18: New and lost joint successes for raw change versus raw clean on all 2,348 quartets.

Seed	New joint successes	Lost joint successes	Net count	ΔJ3 (pp) [95% CI]
11	229	87	142	6.0 [4.0, 8.2]
23	199	105	94	4.0 [1.7, 6.4]
47	166	99	67	2.9 [0.9, 4.8]

On the fresh bank, frequent incomplete repair coexisted with positive net improvement for raw-coordinate models. The near-cancellation of gains and losses on the development bank did not recur. These representation contrasts apply to the tested models; broader interaction effects and held-out model-selection utility remain unestablished.

Separate archived comparison. On the historical 509-quartet bank, geometric-change joint scores were 0.473, 0.448 and 0.436, versus 0.059, 0.047 and 0.041 for raw-clean. The archive contains only these two-arm summaries, without an eligible-parent count. It therefore cannot support parent-cluster intervals or a four-arm interaction estimate.

A.10 S10. Focal Distillation and the repair–preservation trade-off

Objective and paired baseline. The logit-matching implementation follows Focal Distillation from Yan et al. (2021), with alpha=1, beta=5 and binary logits represented as [0,z]. Clean teachers remain frozen. Twelve full-batch runs cross three seeds with lambda=0, 0.01, 0.1 and 1. Within each seed, all runs share initialization, normalization, the 512 clean and 1,534 mined single-edit examples, and 300 optimizer steps. The supervised objective is clean-pool mean BCE plus edit-pool mean BCE. The distillation term averages over the concatenated pools. This task-specific recipe is not a replication of the original ImageNet experiment.

Lambda=1 was fixed as primary; 0.01 and 0.1 are prespecified sensitivity settings. Comparisons use the simultaneously retrained lambda=0 arm, not the historical raw-change checkpoint. The fresh bank was generated after these choices and all checkpoints were fixed. The twelve fits consumed 3,600 optimizer steps and 7,365,600 student-state presentations; cached teacher evaluation added 6,138 states. The additional teacher evaluations make FLOPs unequal despite equal optimizer steps.

Table S19: Primary distillation-minus-paired-baseline contrasts, in percentage points with 95% parent-cluster intervals.

Seed	ΔAB	ΔJ3	Δfull repair / H	Δthree-state negative flips
11	-15.4 [-19.8, -11.1]	-5.0 [-7.0, -3.0]	-8.3 [-11.1, -5.6]	-13.7 [-15.9, -11.7]
23	-11.5 [-15.7, -7.3]	-6.4 [-8.9, -4.0]	-8.5 [-11.3, -5.7]	-13.0 [-14.9, -11.1]
47	-11.2 [-15.7, -6.9]	-4.3 [-6.4, -2.1]	-7.2 [-9.9, -4.7]	-14.3 [-16.3, -12.3]

Negative flips here are the fraction of A/B/AB state rows that change from correct to incorrect relative to the fixed raw-clean model. They are not a rate over ordinary parent scenes, nor a conditional rate among previously correct rows. H also remains the same historical raw-clean failure set within each seed. The primary recipe reduced regression, endpoint accuracy, and full repair, failing to improve repair and preservation together.

Table S20: All distillation settings on the same fresh bank; rates are percentages.

Seed	Lambda	AB	J3	Full repair / H	Three-state negative flips
11	0	60.5	9.5	10.9	18.2
11	0.01	48.2	7.7	7.2	11.1
11	0.1	47.4	5.5	5.0	5.2
11	1	45.1	4.4	2.6	4.4
23	0	55.7	11.1	11.5	16.8
23	0.01	50.6	8.0	9.0	12.1
23	0.1	45.5	5.4	4.1	4.9
23	1	44.2	4.7	3.0	3.8
47	0	54.6	8.9	9.0	17.9
47	0.01	51.0	7.8	7.7	13.3
47	0.1	43.9	5.7	4.2	4.4
47	1	43.4	4.6	1.8	3.6

The sensitivity settings did not exceed their paired baseline's full-repair point estimate, but some contrast intervals include zero. All settings and intervals remain available in the data supplement; no setting was selected post-test. Retrained lambda=0 models differ from the archived raw-change models despite matching training rows and normalization. The two baseline families are reported separately because the cause of their numerical difference was not isolated. The observed trade-off applies to this implementation and does not establish that distillation is generally ineffective.

A.11 Reproduction resources and limitations

The supplementary bundle contains prediction tables, complete correctness transitions, metrics and intervals, figure inputs, generator and analysis code, and configurations. It includes the recorded tracking snapshot and fresh-bank outputs; the source directory regenerates the figures from these inputs. Reconstruction commands recover the principal counts from saved predictions, and bundled coordinate checkpoints and scene banks support model re-inference. Full visual reproduction additionally requires the unbundled visual checkpoints or pretrained initialization and the recorded compute environment. The asset manifest identifies included files and assets held locally without a public download. A permanent external archive is not yet available.

The reported results include fresh-bank evaluation, twelve Focal Distillation runs, and independent replay of the archived structural comparison. Fresh-bank model-selection utility, new structural-model training, and transfer across semantic tasks or renderers were not evaluated for this manuscript. The tracking example illustrates geometric-oracle behavior only; it supplies no trained football-model repair results or pass-success predictions.