GRADIENT COHERENCE IS NOT ENOUGH: SEPARATING INFLUENCE FROM UTILITY IN CONTINUAL TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual test-time adaptation (CTTA) updates a model on unlabeled test data, making it difficult to determine which gradient directions are useful for adaptation. Recent approaches exploit consistency in gradient history to identify structured update directions, based on the intuition that gradients sharing consistent structure provide more reliable adaptation signals. But does a consistent gradient direction actually lead to better predictions? We study this question through a mechanistic analysis of gradient-based CTTA, separating three properties of an update direction: its coherence with past gradients, its influence on model behavior, and its utility for the task. We first establish that gradient coherence is not merely a geometric property: differences in gradient structure propagate through the tracked subspace to actual updates, leading the model toward substantially different parameter and prediction trajectories. Having established that coherent gradient structure can meaningfully influence adaptation, we then ask whether this influence translates into task improvement by comparing candidate directions from identical model states. Across five independently permuted trajectories and 15 frozen states, the more coherent direction in a pair is also the more influential one in 68.8% of comparisons. In contrast, the more coherent direction provides greater supervised improvement in only 47.5% of comparisons, showing no corresponding positive utility-ranking pattern. These results show that gradient coherence captures real influence over adaptation, but influence is not the same as utility: a consistent direction can strongly change the model without being a better direction for improving its predictions.

1 INTRODUCTION

In continual test-time adaptation (CTTA), a pretrained model is updated using an unlabeled data stream under changing test conditions. Because labels are unavailable during adaptation, updates are guided by objectives that do not directly measure prediction correctness. Their reliability becomes particularly important when the incoming distribution changes, as differences between successive batches can produce conflicting optimization objectives during online adaptation (Wang et al., 2024). When individual batches provide conflicting signals, consistency across adaptation steps offers a possible basis for guiding updates. Its value for adaptation, however, depends on whether the directions identified through this consistency lead to better predictions. The relationship between a consistent optimization signal and task improvement therefore deserves direct examination.

Consistent update directions can be identified by examining gradients collected over successive adaptation steps. Through this history, components shared across batches can be distinguished from variations associated with individual batches, providing a basis for using gradient structure to guide adaptation. In LCoTTA, this approach is motivated by the observation that gradients associated with entropy-truthful samples are more correlated than those associated with entropy-deceptive samples. A principal subspace is tracked from gradient history, and gradients are projected onto this subspace before the model is updated (Duan et al., 2025). This mechanism provides an explicit connection between historical gradient structure and parameter updates. It allows a separate analytical question to be examined: whether compatibility with that structure ranks candidate directions by their predictive influence and supervised utility.

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.

To examine this relationship, three properties of an update direction are distinguished: its *coherence* with gradient structure, its *influence* on model behavior, and its *utility* for the task. Coherence describes the relationship between a direction and the gradient structure used as a reference. Influence is measured by the change in model predictions, whereas utility is measured by the improvement achieved on the task. A change in the predicted probability distribution indicates influence, but its value for the task depends on whether probability is assigned more accurately to the correct class. Since predictions can change substantially without becoming more accurate, greater influence cannot by itself be interpreted as greater utility. This distinction is also difficult to establish from an algorithm's overall performance, which reflects the combined effects of several components over an evolving sequence of model states. A direct comparison is therefore needed: **at the same model state, is a more coherent direction more influential, and is it also more useful?**

This question is investigated through controlled comparisons of gradient directions, using the historical-subspace mechanism implemented by LCoTTA (Duan et al., 2025) as the experimental setting. Its explicit connection between gradient history, projection, and optimization allows structural changes to be traced through updates and predictions. Gradient computation is first varied under controlled conditions to determine whether differences in structure are reflected in the subspace, updates, and subsequent model behavior. As the resulting adaptation trajectories diverge, differences in their model states make the quality of individual directions difficult to compare directly. A complementary evaluation is therefore conducted by restoring the same full model state for each candidate direction, applying perturbations with equal norms, and measuring their effects on a common evaluation set. For each frozen state, coherence is assessed against a subspace estimated from gradients collected before that state. The same evaluation batches are retained across candidates so that prediction changes are measured under the same evaluation conditions. Influence is assessed through changes in predictive distributions, while utility is assessed through reductions in supervised cross-entropy. Labels are used only to evaluate utility in this comparison; they are not used to construct or select candidate directions. In this way, the two properties can be examined while the starting state, perturbation magnitude, and evaluation data are held fixed.

The results show that differences in gradient computation are reflected in the tracked subspaces and projected gradients, leading to substantially different parameter and prediction trajectories. The same-state comparison then examines separately how coherence ranks predictive influence and supervised improvement. In the ResNet-50 study on ImageNet-C (Hendrycks & Dietterich, 2019), directions are compared at 15 frozen states obtained from five independently permuted trajectories. After averaging within-state pairwise concordance across states and trajectories, the more coherent direction is also the more influential one in **68.8%** of eligible comparisons, and provides greater supervised improvement in **47.5%**. A similar contrast is observed in a vision transformer (ViT) adaptation setting evaluated on two target corruptions: influence is consistently ranked by coherence, while utility ranking is weaker and varies across states and continuations. This extends the observation to another adaptation setting, although the effects of architecture cannot be isolated because several configuration choices are changed together. These comparisons concern the supervised improvement produced by a single perturbation on a fixed evaluation set, rather than the final accuracy reached after continued adaptation. Across the tested settings, coherence is therefore associated with real influence on adaptation, but this association does not provide equivalent evidence of task improvement. These findings support a distinction between directions that strongly change predictions and directions that improve them, with the following contributions:

- **Mechanistic evidence:** The dependence of gradient geometry on its computation is characterized, and its effects are traced through tracked subspaces to projected gradients and adaptation trajectories. This establishes how differences in gradient structure affect model behavior during adaptation.
- **Controlled directional evaluation:** A comparison is developed in which coherence, predictive influence, and supervised utility are measured separately for candidate directions. Their starting state, perturbation norm, and evaluation data are held fixed so that differences between directions can be assessed.
- **Empirical separation of influence and utility:** Consistent influence ranking is observed without a corresponding positive utility-ranking pattern in the primary ResNet study. The contrast recurs under additional model and corruption settings within the studied adaptation mechanism.

2 Related Work

Continual adaptation under changing test conditions: Continual test-time adaptation requires a pretrained model to accommodate changing distributions while limiting error accumulation and forgetting. These challenges have been addressed through complementary changes to the model, adaptation objective, and use of incoming data. In BECoTTA (Lee et al., 2024a), domain-dependent adaptation is supported by routing inputs through low-rank experts. At the objective level, conflicts between successive batches are addressed through distribution alignment (Wang et al., 2024), while changing environments are accommodated by combining representation alignment over different lengths of stream history in Ada-ReAlign (Zhang et al., 2024). Reliable information can also be retained through uncertainty-aware buffering and replay (Zhu et al., 2024). The effectiveness of adaptation further depends on how test data are normalized: unreliable statistics from small batches are addressed by MemBN (Kang et al., 2024), whereas statistics distorted by temporal label correlation are addressed by UnMix-TNS (Tomar et al., 2024). These studies motivate attention to the conditions under which adaptation signals are computed, alongside the mechanisms through which they affect model updates. The normalization methods address the reliability of statistics used for prediction and adaptation. They provide relevant computational context, but do not directly test how historical gradient coherence ranks the utility of candidate directions.

Gradient structure and parameter updates: Gradient structure has been used more directly to determine which updates should be applied. In GALA (Sahoo et al., 2025), layers are selected using the alignment between a proposed update and the anticipated cumulative parameter displacement. Because this criterion can restrict necessary changes in direction, its reference parameters are periodically reset. A different form of alignment is considered in CoCo-MT-TTA (Ye et al., 2026), where consensus between task gradients is combined with constraints on parameter changes. In LCoTTA (Duan et al., 2025), correlations among entropy gradients motivate adaptation within a principal subspace tracked from gradient history. These methods use distinct notions of alignment: agreement with accumulated updates, agreement across tasks, and compatibility with a historical subspace. Their interpretation is complemented by studies of parameter dynamics. Bayesian filtering is used to track latent weights in SLWI (Lee & Chang, 2024), while time-varying covariance associated with gradient noise is incorporated into steady-state adaptation (Lee, 2025a). Parameter evolution is also modeled through stochastic differential equations in SIGMA (Lee, 2025b). Together, these studies connect update structure and parameter evolution with adaptation performance. An improvement obtained by a complete adaptation method does not, by itself, establish how its structural signal ranks alternative directions when the remaining components and the starting model state are held fixed. Our analysis examines a more specific relationship within this setting: whether compatibility with historical gradient structure ranks candidate directions by their influence on predictions and by their utility for the task. LCoTTA supplies the historical-subspace mechanism used to examine this relationship under controlled conditions. The different alignment criteria above are not treated as interchangeable.

Reliability of adaptation signals: The reliability of unlabeled adaptation signals has also been examined through model confidence and predictive performance. DeYO (Lee et al., 2024b) identifies limitations of entropy-based confidence under biased conditions and supplements it with prediction changes induced by object-destructive transformations. Related concerns about overconfidence are addressed through conservative entropy minimization in COME (Zhang et al., 2025). Beyond confidence control, ReTTA (Park et al., 2025) analyzes the relationship between entropy and energy, showing that reducing the former does not ensure reduction of the latter. An alternative use of entropy is developed by Bar et al. (2024), who combine shift detection with matching source and test entropy distributions. At the model level, Kim et al. (2024) study how adaptation strengthens relationships between prediction agreement and accuracy, supporting performance estimation and model selection without target labels. Schirmer et al. (2025) develop sequential risk monitoring for adapting models using unlabeled loss proxies under specified assumptions. These studies assess sample reliability or model performance. Our study addresses the interpretation of gradient coherence at the level of individual update directions, comparing candidates from identical model states under matched conditions. In this comparison, each candidate is applied to the same restored state with an equal perturbation norm, and predictions are evaluated on the same data. Supervised labels are used only to assess utility. By measuring behavioral change separately from supervised improvement, the analysis distinguishes evidence that a direction influences adaptation from evidence that it improves the task.

3 Study Framework and Experimental Design

3.1 Analytical Setting

Two complementary comparisons examine gradient structure and adaptation. Controlled interventions test whether changes in gradient computation alter geometry and model behavior. Directions are then compared from identical adaptation states to determine whether greater coherence indicates greater influence and utility.

LCoTTA (Duan et al., 2025) provides an experimentally accessible historical-subspace mechanism: gradient history determines the projector applied before the optimizer step. This connection allows the effects of structural interventions to be traced through updates and predictions (Figure 1a). Only normalization-affine parameters are adapted; other learned parameters remain fixed.

3.2 Coherence, Influence, and Utility

Three properties are distinguished for each candidate direction j. **Coherence** C_j measures how much of its gradient is retained by the historical subspace. **Influence** I_j measures how much predictions change after a perturbation. **Utility** U_j measures whether that change reduces supervised cross-entropy.

Each candidate is evaluated from the same reference state s, which includes the adapted parameters $\boldsymbol{\theta}$, model buffers, optimizer state, and adaptation history. Let $\mathbf{g}_j$ denote the gradient of the specified adaptation objective with respect to $\boldsymbol{\theta}$, computed for candidate batch j at this state. Let $\mathbf{P}$ denote the orthogonal projector onto the subspace estimated from preceding gradients. For a finite, nonzero candidate gradient, coherence is defined as the fraction of its squared norm retained after projection:

$$C_j = \frac{\|\mathbf{P}\mathbf{g}_j\|_2^2}{\|\mathbf{g}_j\|_2^2}. \tag{1}$$

To evaluate the candidate's effect, a perturbation $\mathbf{d}_j$ is applied to the adapted parameters. The resulting state s_j is initialized from s, with $\boldsymbol{\theta}$ replaced by $\boldsymbol{\theta} + \mathbf{d}_j$. Predictions before and after the perturbation are compared on the same evaluation probe $\mathcal{E}$, with its image identities, order, and batch partition held fixed; $|\mathcal{E}|$ counts images. The predictive distribution at state s is denoted by p_s. Influence is measured by the forward Kullback–Leibler (KL) divergence from reference to perturbed predictions. Utility is measured using mean supervised cross-entropy, denoted by $\mathcal{L}_{\mathrm{CE}}$, with labels accessed only for this evaluation:

$$I_j = \frac{1}{|\mathcal{E}|} \sum_{x \in \mathcal{E}} \mathrm{KL}\big(p_s(\cdot \mid x) \,\|\, p_{s_j}(\cdot \mid x)\big)\,, \tag{2}$$

$$U_j = \mathcal{L}_{\mathrm{CE}}(s; \mathcal{E}) - \mathcal{L}_{\mathrm{CE}}(s_j; \mathcal{E}). \tag{3}$$

A larger I_j indicates a greater change in predictions, whereas a positive U_j indicates lower supervised loss. Section 6 describes how perturbation norms are matched and candidate directions are ranked. The earlier geometry analysis uses pairwise gradient cosine to measure agreement between gradients; this differs from the historical-subspace score C_j.

3.3 Experimental Regimes and Controls

On ImageNet-C (Hendrycks & Dietterich, 2019) severity 5, initial ResNet-50 diagnostics use mean-entropy gradients with label-informed grouping and ordering. Subsequent adaptation interventions and directional comparisons retain the studied update mechanism on label-blind permutations. Directional evaluation also uses ViT-B/16 after target-excluded warm-up. Objectives, streams, and settings are detailed in Appendix A.

Computation interventions hold state and batch fixed. Directional comparisons restore the same full state for each candidate, apply equal-norm perturbations directly without optimizer steps, and retain the same probe and batch grouping (Figure 1b). In the directional comparisons, labels enter only utility evaluation. The ViT comparison tests another regime of the same historical-subspace mechanism, without isolating architecture from other configuration changes.

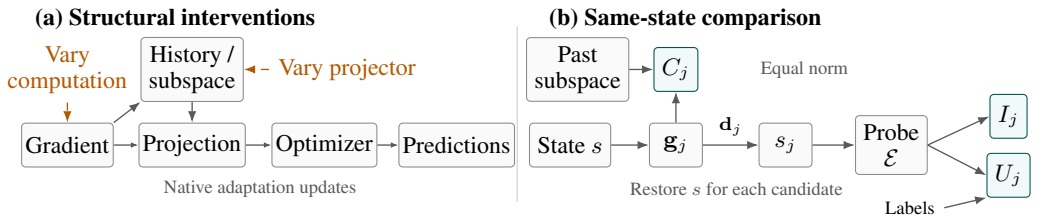

Figure 1: Controlled interventions and same-state directional comparisons separating gradient coherence, predictive influence, and supervised utility.

4 Gradient Geometry Depends on Its Computation

4.1 Coherence under Controlled Computation Changes

Following Duan et al. (2025), images are **entropy-truthful (ET)** when the source model assigns the true class a strictly higher probability than every alternative, and **entropy-deceptive (ED)** otherwise, including ties. These source-defined groups provide a fixed diagnostic partition for testing computation sensitivity; their membership describes prediction correctness rather than update utility.

ResNet-50 is adapted by mean batch entropy on severity-5 ImageNet-C Gaussian Noise, ordered by source probability margin. At each pre-update state, gradients are computed on the same batch using current batch statistics (**BATCH**) or pretrained running statistics (**FROZEN**). Only BATCH advances the trajectory.

Mean pairwise gradient cosine, distinct from C_j, is compared between nine ET gradients and 1,000 size-nine subsets of 247 ED gradients in the first 256 positions. The subsets overlap and are not independent replications. Under computation condition b, the gap is

$$\Delta^{b}_{\text{gap}} = \overline{\cos}^{b}_{\text{ET}} - \overline{\cos}^{b}_{\text{ED,matched}}. \tag{4}$$

With membership fixed, FROZEN increases both groups' coherence and changes the ET–ED gap from $+0.0182$ to -0.0222 (Figure 2a). The measured gap therefore depends on the normalization computation. Since ET and ED occupy different stream regions, this comparison does not isolate correctness from adaptation time.

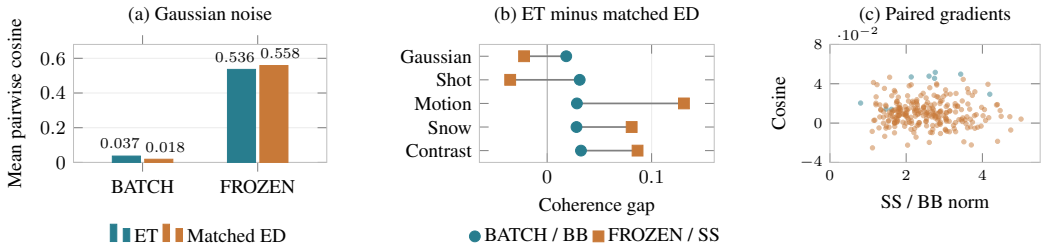

Figure 2: Normalization statistics change absolute coherence and the ET–ED gap, with different effects across corruption conditions.

4.2 Changes in Direction and Magnitude

The paired observations in Figure 2(c) have mean BATCH–FROZEN cosine 0.0110 and mean FROZEN-to-BATCH norm ratio 2.525. Direction and magnitude both change, ruling out rescaling alone. Appendix B examines shared directional structure and interactions between normalization means and variances.

Figure 2(b) shows that the gap decreases under shot noise but increases under motion blur, snow, and contrast; BB and SS denote BATCH and FROZEN. Sensitivity recurs with varying direction, without establishing better update utility. Section 5 examines its effect on adaptation.

5 Gradient Geometry Influences Adaptation

5.1 From Tracked Subspaces to Projected Gradients

The operational effect of computation-dependent geometry is tested by intervening on the historical projector applied to a fixed gradient. Reference adaptation follows the LCoTTA mechanism on a label-free permutation of 50,000 Gaussian Noise images. At 20 frozen anchors, the same raw BB gradient is projected through subspaces estimated from BB or SS histories.

Full SS uses its own sample selection and entropy weights; selection-fixed SS retains BB-selected samples and weights for its historical gradients. Let $\mathbf{P}_a^{\mathrm{BB}}$ and $\mathbf{P}_a^{\mathrm{SS,fixed}}$ denote the corresponding projectors at anchor a, and $\mathbf{g}_a^{\mathrm{BB}}$ the shared raw gradient. Their directional discrepancy is

$$D_{\mathrm{proj},a}^{\mathrm{fixed}} = 1 - \cos\left(\mathbf{P}_a^{\mathrm{BB}}\mathbf{g}_a^{\mathrm{BB}}, \mathbf{P}_a^{\mathrm{SS,fixed}}\mathbf{g}_a^{\mathrm{BB}}\right). \tag{5}$$

The full intervention is evaluated analogously.

Figure 3(a) shows median discrepancies of 0.9484 for full and 0.9547 for selection-fixed interventions, compared with approximately zero for BB replay. Both exceed replay at all 20 anchors, so selection and weighting changes do not fully explain the effect. Random subspaces also produce large discrepancies, providing context for the magnitude of directional change. Section 6 separately examines utility through same-state comparisons. These measurements concern projected gradients, not optimizer displacements; Appendix E provides magnitude comparisons and controls.

5.2 From Projected Gradients to Predictions

Two branches evolve from the same complete state using BB-derived projectors (Arm A) or selection-fixed SS-derived projectors (Arm B). Both retain BB adaptation computations, with independently evolving gradients and histories. Let $\boldsymbol{\theta}_t^{\mathrm{A}}$ and $\boldsymbol{\theta}_t^{\mathrm{B}}$ denote the adapted normalization-affine vectors after update t in the two branches. Figure 3(c) shows their Euclidean distance, defined in Eq. (16).

Across starts at batches 1300, 1900, and 2500, predictions disagree on 8.45–10.43% of suffix images, whereas net accuracy differences range from $+0.036$ to $+0.178$ percentage points. The accuracy differences remain below the preregistered one-point materiality threshold. Figure 3(b) explains this contrast: 75.9–77.5% of disagreements involve different predictions that are both wrong, while opposing correctness changes largely cancel.

The intervention therefore separates substantial changes in predictions from smaller net changes in correctness. This does not establish statistical equivalence. The three starts have overlapping suffixes and are not independent replications. Moreover, diverging trajectories do not directly compare direction quality at a shared state. Section 6 addresses that comparison.

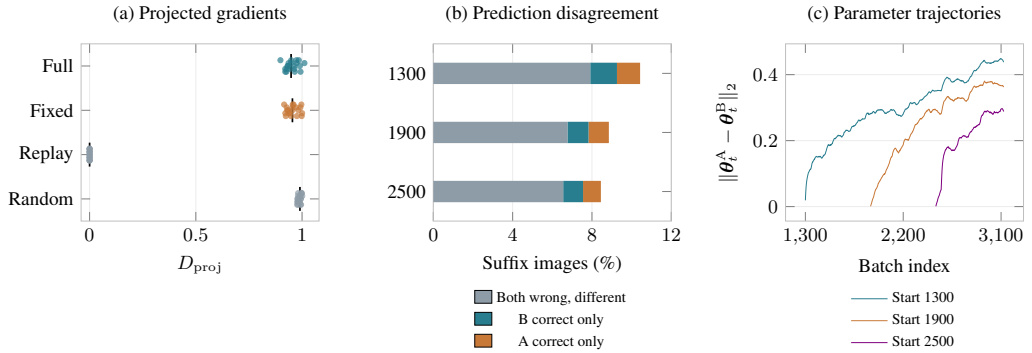

Figure 3: Projector interventions change gradients, parameter trajectories, and predictions; opposing correctness transitions yield smaller net accuracy differences.

6 SEPARATING INFLUENCE FROM UTILITY

6.1 COMPARING DIRECTIONS AT THE SAME STATE

The predictive value of historical coherence is evaluated by comparing original candidate gradient directions from the same complete adaptation state, with equal perturbation norms and a common probe (Table 1). For trajectory r and anchor a, state s_{ra} is frozen before processing the anchor batch. At this state, the quantities defined in Section 3 are indexed by ra and by j for each candidate, while $\boldsymbol{\theta}$ denotes the adapted parameter vector of s_{ra}. The projector $\mathbf{P}_{ra}$ uses strictly preceding gradient snapshots. Each candidate gradient $\mathbf{g}_{raj}$ is computed from the restored state using the native objective and is excluded from the reference history.

To separate direction from gradient magnitude, each candidate is applied as

$$\mathbf{d}_{raj} = -\delta_{ra} \frac{\mathbf{g}_{raj}}{\|\mathbf{g}_{raj}\|_2}. \tag{6}$$

Here, δ_{ra} is the native anchor-update displacement norm, measured on an isolated state copy. Each candidate branch restores the reference state and replaces the adapted parameters with $\boldsymbol{\theta} + \mathbf{d}_{raj}$ directly, without running the native optimizer.

Coherence, influence, and utility follow Section 3, using the same probe $\mathcal{E}_{ra}$, order, and batch partition. Labels enter only one-step cross-entropy utility evaluation, which does not measure long-horizon accuracy.

The projector is used to measure coherence, while each perturbation follows the original candidate gradient direction. The comparison therefore evaluates how compatibility with historical structure ranks these directions by predictive influence and one-step supervised improvement. The disjoint evaluation probe contains the 64 local-stream batches immediately following the 32 candidate batches. This is an offline comparison on future local data, not an online selection procedure. ResNet-50 uses current-batch normalization, making the fixed probe partition an evaluation control. Probe-induced state changes are discarded before assessing another candidate.

Table 1: Directional comparisons use matched conditions within each anchor.

Design component	ResNet-50	ViT
Targets	Gaussian noise	Gaussian noise; motion blur
Runs $\times$ anchors $\times$ candidates, per target	$5 \times 3 \times 32$	$5 \times 3 \times 32$
Probe images / batch size	1,024 / 16	1,024 / 16

6.2 COHERENCE RANKS INFLUENCE MORE CONSISTENTLY THAN UTILITY

Ranking is evaluated within each anchor. A candidate pair is concordant when the direction with greater coherence also has greater influence or utility. Pairs whose differences fall within fixed numerical tie tolerances are excluded from the corresponding comparison.

Let $\mathcal{J}_{ra}^{CU}$ contain the eligible pairs of candidates indexed by j and k, represented once by $j < k$. Utility concordance is defined as

$$q_{ra}^{CU} = \frac{\#\left\{(j,k) \in \mathcal{J}_{ra}^{CU} : (C_{raj} - C_{rak})(U_{raj} - U_{rak}) > 0\right\}}{|\mathcal{J}_{ra}^{CU}|}. \tag{7}$$

Influence concordance q_{ra}^{CI} is defined analogously. C–U concordance is the primary endpoint, while C–I is the secondary comparison. Scores are averaged equally over the three anchors within each trajectory and then over the five trajectories. Candidate pairs are not treated as independent replications.

For ResNet-50, aggregate concordance is **68.8% for C–I** and **47.5% for C–U**. Influence concordance exceeds 0.5 at all 15 anchors and utility concordance at five. Figure 4(a) shows the paired scores at each anchor, with colors identifying trajectories and black bars marking aggregate means.

These results support a consistent positive influence-ranking pattern without a corresponding positive utility-ranking pattern. The distinction concerns relative ordering: a concordant C–U pair identifies the direction with greater supervised improvement, but does not imply that either direction has positive utility. Likewise, a C–U score near 0.5 does not establish statistical equivalence to random ranking or independence between coherence and utility.

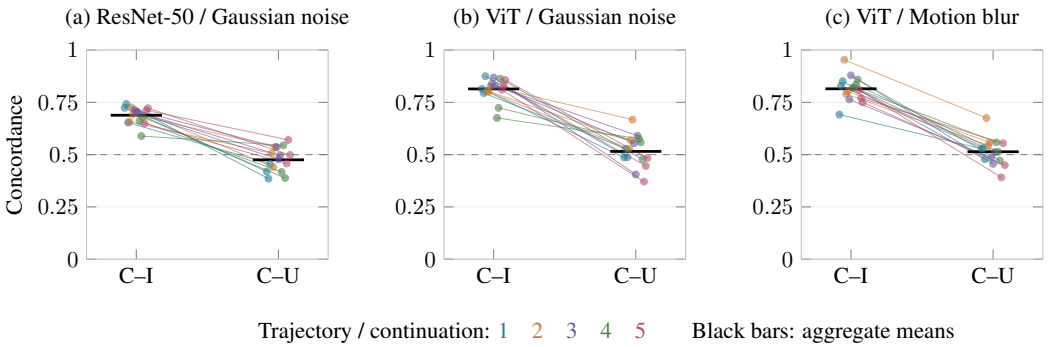

Figure 4: Historical coherence ranks predictive influence more consistently than utility across the three tested adaptation settings.

6.3 The Contrast Persists in a ViT Adaptation Regime

The same comparison is applied within the historical-subspace mechanism using a ViT-B/16 model with LayerNorm-affine updates. Gaussian noise and motion blur are evaluated separately. For each target, five independently permuted continuations share a deterministic target-excluded warm-up. Three anchors per continuation give 30 states and 960 candidate directions across both targets.

Aggregate C–I concordance is **81.4% for Gaussian noise** and **81.5% for motion blur**, with scores above 0.5 at all 30 anchors. Corresponding C–U concordance is **51.5%** and **51.4%**, exceeding 0.5 at eight and nine of the 15 anchors, respectively. Figures 4(b–c) show variable utility ranking alongside consistently positive influence ranking.

The slightly positive C–U scores do not reproduce the negative ResNet-50 sign; stronger and more consistent influence ranking recurs. Architecture, normalization, learning rate, subspace settings, and preceding history change together, so quantitative differences cannot be attributed to architecture alone.

Consistency across summaries and runs: Table 2 examines the contrast using within-anchor Spearman correlations, averaged equally over anchors and runs. These also show a stronger coherence–influence relationship. Utility concordance also varies across runs: its mean exceeds 0.5 in two of five ResNet-50 trajectories and four and three ViT continuations, respectively. Thus, positive ViT averages do not describe every continuation. These summaries are neither independent evidence nor significance tests; ViT continuations remain conditional on their shared warm-up.

Table 2: Secondary associations and run-level consistency distinguish influence ranking from utility ranking.

Setting	Mean within-anchor Spearman C–I	C–U	Runs with mean C–U concordance > 0.5
ResNet-50 / Gaussian noise	0.517	−0.082	2/5
ViT / Gaussian noise	0.798	0.042	4/5
ViT / Motion blur	0.799	0.034	3/5

7 DISCUSSION

The findings clarify what gradient coherence can indicate in continual test-time adaptation. The measured geometry changes with the gradient computation, and differences in the tracked subspace propagate to projected gradients, parameter trajectories, and predictions. Coherence therefore captures structure that can meaningfully affect adaptation. However, these effects do not by themselves establish that the resulting updates improve predictions.

The same-state comparison separates these questions by holding the reference state, perturbation norm, and evaluation probe fixed. Under these controls, coherence ranks predictive influence more consistently than supervised improvement. The contrast also appears in the ViT adaptation settings, although their aggregate utility concordance is slightly above 0.5. The shared finding is therefore stronger and more consistent influence ranking, rather than uniformly negative utility ranking.

The analytical contribution is to distinguish the information that historical coherence provides about predictive influence from the information it provides about supervised improvement. Evaluating these relationships under matched conditions clarifies how a structural adaptation signal can be interpreted within the tested regimes.

8 LIMITATIONS

The empirical tests use the historical-subspace mechanism implemented by LCoTTA across selected models and corruption settings. The resulting evidence is limited to these regimes and does not establish a general relationship across datasets or adaptation-method families. The ViT extension changes architecture, normalization, optimization, and subspace settings together, so quantitative differences cannot be attributed to any one factor.

Directional utility is measured through one-step, equal-norm perturbations and supervised cross-entropy on a finite probe. It does not establish long-horizon accuracy or behavior at other step sizes. The early geometry diagnostics use label-informed grouping and ordering; labels serve diagnostic purposes rather than an available online selection rule. Pairwise cosine, cross-fitted retention, and historical-subspace coherence also measure different aspects of gradient structure.

Dependence remains within the evaluation design: anchors share trajectories, trajectory comparisons have overlapping suffixes, and ViT continuations share a warm-up within each target. Candidate pairs are therefore not independent replications. The results establish neither statistical equivalence nor independence between coherence and utility, and they do not identify a better direction-selection criterion.

9 CONCLUSION

Does a gradient direction that agrees with past gradients provide a better adaptation signal? This paper examined that question by distinguishing coherence, predictive influence, and task utility in continual test-time adaptation. The analysis first showed that gradient geometry depends on its computation and that differences in the tracked subspace propagate to actual updates, parameter trajectories, and predictions. These observations establish that the measured structure can meaningfully affect adaptation. Whether that influence improves predictions was then examined by comparing candidate directions from identical model states, with equal perturbation norms and a common evaluation probe. Mean within-state concordance, averaged over five independently permuted trajectories and 15 frozen states, was 68.8% for predictive influence and 47.5% for supervised improvement. The contrast between stronger influence ranking and weaker utility ranking recurred in the additional ViT adaptation settings. Within the tested settings, historical coherence ranks predictive influence more consistently than supervised improvement. The analysis therefore clarifies the information supplied by this structural signal and supports evaluating its relationships with influence and utility separately under controlled conditions.

AI USE STATEMENT

AI tools supported writing and polishing, literature retrieval and discovery, and research ideation and execution, including implementation and analysis. The authors take full responsibility for the manuscript, results, and conclusions.

REFERENCES

Yarin Bar, Shalev Shaer, and Yaniv Romano. Protected test-time adaptation via online entropy matching: A betting approach. In *Advances in Neural Information Processing Systems*, volume 37, pp. 85467–85499. Curran Associates, Inc., 2024. doi: 10.52202/079017-2714. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9b35a0a20d617dc68ae98a7a57df2f51-Abstract-Conference.html.

Dexin Duan, Rui Xu, Peilin Liu, and Fei Wen. Lifelong test-time adaptation via online learning in tracked low-dimensional subspace. In *Advances in Neural Information Processing Systems*, volume 38, pp. 19024–19053, 2025. doi: 10.52202/085713-0640. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1b6f528fe522a956dce80d5b98be8cd5-Abstract-Conference.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Juwon Kang, Nayeong Kim, Jungseul Ok, and Suha Kwak. MemBN: Robust test-time adaptation via batch norm with statistics memory. In *Computer Vision – ECCV 2024*, pp. 467–483. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73390-1_27. URL https://doi.org/10.1007/978-3-031-73390-1_27.

Eungyeup Kim, Mingjie Sun, Christina Baek, Aditi Raghunathan, and J. Zico Kolter. Test-time adaptation induces stronger accuracy and agreement-on-the-line. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120184–120220. Curran Associates, Inc., 2024. doi: 10.52202/079017-3820. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d96fcc07d623a9eba68616629911143a-Abstract-Conference.html.

Daeun Lee, Jaehong Yoon, and Sung Ju Hwang. BECoTTA: Input-dependent online blending of experts for continual test-time adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27072–27093. PMLR, 2024a. URL https://proceedings.mlr.press/v235/lee24ab.html.

Jae-Hong Lee. Bayesian weight enhancement with steady-state adaptation for test-time adaptation in dynamic environments. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32889–32906. PMLR, 2025a. URL https://proceedings.mlr.press/v267/lee25b.html.

Jae-Hong Lee. Parameter dynamics of online machine learning and test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 68273–68302. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2294. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/62ab1c2cb4b03e717005479efb211841-Abstract-Conference.html.

Jae-Hong Lee and Joon-Hyuk Chang. Stationary latent weight inference for unreliable observations from online test-time adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26379–26395. PMLR, 2024. URL https://proceedings.mlr.press/v235/lee24b.html.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, pp. 46977–47002, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/cd0665986c1e9c15f6569ac944bcf88a-Abstract-Conference.html.

Mincheol Park, Heeji Won, Won Woo Ro, and Suhyun Kim. Rethinking entropy in test-time adaptation: The missing piece from energy duality. In *Advances in Neural Information Processing Systems*, volume 38, pp. 47123–47140. Curran Associates, Inc., 2025. doi: 10.52202/085713-1570. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/435422305988b73c6cc00bcb29ba2531-Abstract-Conference.html.

Sabyasachi Sahoo, Mostafa ElAraby, Jonas Ngnawe, Yann Batiste Pequignot, Frédéric Precioso, and Christian Gagné. A layer selection approach to test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20237–20245, 2025. doi: 10.1609/aaai.v39i19.34229. URL https://doi.org/10.1609/aaai.v39i19.34229.

Mona Schirmer, Metod Jazbec, Christian Andersson Naesseth, and Eric Nalisnick. Monitoring risks in test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 80915–80948. Curran Associates, Inc., 2025. doi: 10.52202/085713-2705. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/746960ad49ddb47248970a0e1404230c-Abstract-Conference.html.

Devavrat Tomar, Guillaume Vray, Jean-Philippe Thiran, and Behzad Bozorgtabar. Unmixing test-time normalization statistics: Combatting label temporal correlation. In *International Conference on Learning Representations*, pp. 50967–50988, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/de08b3ee7c0043a76ee4a44fe68e90bc-Abstract-Conference.html.

Ziqiang Wang, Zhixiang Chi, Yanan Wu, Li Gu, Zhi Liu, Konstantinos Plataniotis, and Yang Wang. Distribution alignment for fully test-time adaptation with dynamic online data streams. In *Computer Vision – ECCV 2024*, pp. 332–349. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72691-0_19. URL https://doi.org/10.1007/978-3-031-72691-0_19.

Zhong Ye, Yu Hu, and Zhenguo Yang. Multi-task test-time adaptation via gradient consensus and plasticity constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27836–27843, 2026. doi: 10.1609/aaai.v40i33.40006. URL https://doi.org/10.1609/aaai.v40i33.40006.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: Test-time adaption by conservatively minimizing entropy. In *International Conference on Learning Representations*, pp. 18506–18535, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

Zhen-Yu Zhang, Zhiyu Xie, Huaxiu Yao, and Masashi Sugiyama. Test-time adaptation in non-stationary environments via adaptive representation alignment. In *Advances in Neural Information Processing Systems*, volume 37, pp. 94607–94632. Curran Associates, Inc., 2024. doi: 10.52202/079017-2999. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/abe31a12e83111fdf2cfd54deed5a2ce-Abstract-Conference.html.

Zhilin Zhu, Xiaopeng Hong, Zhiheng Ma, Weijun Zhuang, Yaohui Ma, Yong Dai, and Yaowei Wang. Reshaping the online data buffering and organizing mechanism for continual test-time adaptation. In *Computer Vision – ECCV 2024*, pp. 415–433. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73007-8_24. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/10845_ECCV_2024_paper.php.

A Experimental Configurations and Implementation Details

A.1 Models, Data, and Preprocessing

The model and input configurations are summarized in Table 3, where R50 abbreviates ResNet-50 and ViT denotes a vision transformer. Only normalization-affine parameters are adapted, and each model's preprocessing is held fixed across its corresponding comparisons. For ViT, this preprocessing follows the input path of the studied implementation and forms part of the experimental regime evaluated here.

Table 3: Models and input configurations. All corruption conditions use ImageNet-C at severity 5.

Setting	ResNet-50	ViT-B/16
Pretrained model	`Standard_R50`, ImageNet-1K V1 weights	`vit_base_patch16_224`, `augreg2_in21k_ft_in1k` weights
Adapted parameters	BatchNorm scale and shift	LayerNorm scale and shift
Adapted scalar parameters	53,120	38,400
Primary conditions	Gaussian noise	Gaussian noise; motion blur
Additional geometry conditions	Shot noise; motion blur; snow; contrast	Not used
Image conversion	RGB to tensor	RGB to tensor
Additional resize, crop, or augmentation	None in the experimental loader	None in the experimental loader
Channel normalization	Mean $(0.485, 0.456, 0.406)$; standard deviation $(0.229, 0.224, 0.225)$, applied inside the model wrapper	No additional channel normalization in the evaluated input path

A.2 Adaptation Settings

The geometry diagnostics use mean predictive entropy over each batch. The later adaptation and directional analyses retain entropy-based selection, reweighting, and historical-subspace projection. Their principal settings, including stochastic gradient descent (SGD) configurations, are reported in Table 4.

Table 4: Adaptation configurations. Settings apply to the reference trajectories. Equal-norm candidate perturbations are evaluated separately from native optimizer updates.

Setting	R50 geometry diagnostics	R50 adaptation and directional analysis	ViT directional analysis
Batch size	16	16	16
Objective	Mean entropy	Selected, reweighted entropy	Selected, reweighted entropy
Optimizer	SGD	SGD	SGD
Learning rate	1.5×10^{-4}	1.5×10^{-4}	5×10^{-4}
Momentum	0.9	0.9	0.9
Nesterov momentum	Enabled	Enabled	Enabled
Dampening / weight decay	0 / 0	0 / 0	0 / 0
Entropy threshold	Not used	2.763	2.763
Absolute probability-similarity threshold	Not used	0.05	0.05
Running probability coefficient	Not used	0.9	0.9
Historical projection	Disabled	Enabled	Enabled
Subspace rank	Not used	25	50
Maximum history snapshots	Not used	100	100
Snapshot interval, batches	Not used	50	100
Mixed precision	Disabled	Disabled	Disabled

Sample selection and weighting: Samples are first retained when predictive entropy is below the threshold $h_{\max}$. Once a running probability estimate is available, samples must also have absolute

cosine similarity below 0.05 to that estimate. Let $\mathcal{S}_t$ denote the selected sample indices at adaptation step t, x_i the corresponding inputs, and H predictive entropy. For the current state s and a nonempty selected set $\mathcal{S}_t$, the selected, reweighted objective, denoted by the superscript sel, is

$$\mathcal{L}_{\mathrm{ent}}^{\mathrm{sel}} = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \omega_i H(p_s(\cdot \mid x_i)) \,, \tag{8}$$

with detached weights

$$\omega_i = \mathrm{stopgrad}[\exp(h_{\max} - H(p_s(\cdot \mid x_i)))] \,. \tag{9}$$

The operator $\mathrm{stopgrad}$ preserves the weight value while preventing differentiation through the weight. Predictions in these expressions are computed using the current adaptation batch. The running probability estimate is initialized from the first nonempty selected set and subsequently updated with coefficient 0.9 on its previous value. Selection and weighting do not use labels. When selection is empty, the native implementation produces an undefined scalar mean and zero parameter gradients. Its optimizer step is retained, so momentum can still change the parameters, while the running probability estimate remains unchanged.

Historical projection: Stored gradient snapshots form the rows of the history matrix $\mathbf{G}$. Principal component analysis (PCA) is applied to the centered history to obtain a basis $\mathbf{Q}$ with orthonormal columns. The corresponding projector is $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$. For the adaptation gradient $\mathbf{g}_t$ at step t, the projected gradient $\mathbf{P}\mathbf{g}_t$ is passed to the optimizer. Centering is used to estimate the basis; the historical mean is not subtracted from the current gradient before projection. Before sufficient history is available, the unprojected gradient is used.

The native projector analysis follows the tracking convention that can include the current gradient at an anchor. The directional comparison instead constructs its reference projector strictly from preceding snapshots. These conventions serve different comparisons and are retained in their respective controls.

Normalization: ResNet-50 reference trajectories use current-batch BatchNorm statistics without accumulating running estimates. Normalization interventions replace the designated statistics while retaining the same pre-update affine parameters and input batch. Counterfactual gradient computations do not themselves advance an alternative trajectory. Their exact contrasts are specified with the geometry analysis.

A.3 Stream Construction and Evaluation Controls

Stream configurations are summarized in Table 5. Diagnostic ordering uses source-model predictions and labels, whereas subsequent adaptation and directional comparisons use label-blind permutations.

Diagnostic grouping: Let s_{src} denote the source state, x an image, y its true class, and ℓ a class index. Source predictions are evaluated using pretrained running normalization statistics. The source entropy-reliability score (ERS) is

$$\mathrm{ERS}(x, y) = p_{s_{\mathrm{src}}}(y \mid x) - \max_{\ell \neq y} p_{s_{\mathrm{src}}}(\ell \mid x). \tag{10}$$

Positive scores define the entropy-truthful group, and nonpositive scores define the entropy-deceptive group. Samples are sorted by decreasing score and processed in consecutive batches. Group assignments remain fixed across the paired computation comparisons. They describe source-model predictions and do not directly identify whether an update improves the task at a later state. The analyzed gradients are gradients of batch-mean entropy rather than independently computed per-image gradients.

ResNet-50 streams: The projector and trajectory interventions use a label-blind permutation of 50,000 Gaussian-noise images. The principal directional comparison uses five permutations, each initialized from the source state. At batch size 16, each stream contains 3,125 batches. Candidate and probe positions are fixed before supervised utility is evaluated.

Target-excluded warm-up: Each ViT target condition starts from a fresh source model. Its target corruption is removed from the fixed 15-corruption sequence, and the remaining 14 conditions are

Table 5: Stream construction and initialization.

Setting	Geometry diagnostics	R50 directional comparison	ViT directional comparison
Images per condition or target stream	5,000	50,000	50,000
Ordering	Decreasing source entropy-reliability score	Label-blind permutation	Label-blind target permutation
Trajectories / continuations	Canonical stream per condition	Five source-start trajectories	Five continuations per target
Permutation seeds	Not used	1–5	1–5
Pre-target warm-up	None	None	17 target-excluded segments
Warm-up sample appearances	Not used	Not used	85,000
Warm-up batches	Not used	Not used	5,321
Shared initialization	Source state	Source state	Common post-warm-up state within each target
Labels used for ordering/grouping	Yes	No	No

traversed once, followed by their first three conditions again. Each segment contains 5,000 deterministically selected images, with no reset between segments. At batch size 16, a segment contains 312 complete batches and one final batch of eight images. The resulting warm-up contains 85,000 sample appearances over 5,321 batches; repeated segments do not represent additional unique conditions.

Within each target, five independently permuted continuations start from the same post-warm-up state. They therefore share an adaptation history rather than constituting five independent complete histories. Gaussian-noise and motion-blur targets have separate source initializations and warm-ups. The ViT study uses batch size 16 to retain the batch size of the controlled directional comparison; the studied implementation's ViT benchmark uses batch size 64.

Directional controls: Candidate branches restore the same full reference state before their parameter perturbations are applied. The evaluation probe retains identical sample identities, order, and batch partition across candidates. This partition is part of the evaluation convention because BatchNorm predictions can depend on the other samples in the batch.

Candidate construction, gradient computation, coherence C_j, and predictive influence I_j are evaluated without labels. Labels are accessed only for supervised utility U_j. Details of candidate and probe construction, state-restoration checks, and calibrated numerical tolerances are provided in Appendix F.

B Computation-Dependent Gradient Geometry

B.1 Diagnostic Groups and Matched Comparisons

Diagnostic membership is determined by the source probability margin defined in Eq. (10). The ET and ED groups follow the criteria specified in Appendix A.3 and provide a fixed partition for examining gradient geometry under paired computation conditions. Labels determine the diagnostic grouping and ordering, while adaptation remains governed by mean predictive entropy. Table 6 summarizes the shared experimental settings, with optimizer and model configurations provided in Appendix A.

The diagnostic stream is ordered by the source probability margin. Each observation is a batch gradient with respect to the adapted normalization-affine parameters $\boldsymbol{\theta}$, rather than an individual-image gradient. Consequently, the ET–ED comparison also reflects the different trajectory regions occupied by the groups. It does not isolate correctness from adaptation time. Only batches containing a single group enter the group summaries; mixed batches remain in the adaptation stream.

Table 6: Settings for the primary gradient-geometry diagnostic.

Setting	Specification
Model	Standard ResNet-50
Primary corruption	ImageNet-C Gaussian Noise, severity 5
Images / batch size	5,000 / 16
Adapted parameters	BatchNorm affine parameters
Diagnostic objective	Mean batch entropy
Primary analysis window	First 256 gradient positions
Group sizes in this window	9 ET and 247 ED gradients
Matched comparison	1,000 ED subsets, each containing 9 gradients

Within each group or subset, coherence is calculated by averaging cosine similarity over all distinct gradient pairs. Let $\mathbf{g}_n^b$ and $\mathbf{g}_m^b$ denote the nonzero batch-gradient observations indexed by n and m under computation condition b. Their pairwise cosine is

$$\cos(\mathbf{g}_n^b, \mathbf{g}_m^b) = \frac{(\mathbf{g}_n^b)^\top \mathbf{g}_m^b}{\|\mathbf{g}_n^b\|_2 \|\mathbf{g}_m^b\|_2}. \tag{11}$$

The ET group average is denoted by $\overline{\cos}^b_{\mathrm{ET}}$. Each ED subset is sampled without replacement, but subsets may overlap. The matched ED value, $\overline{\cos}^b_{\mathrm{ED,matched}}$, is the mean of the 1,000 subset-level values. Their variation describes sensitivity to subset selection within the observed pool, not uncertainty across independent adaptation trajectories.

In the initial diagnostic, ET coherence was 0.036406, compared with a matched ED mean of 0.018473, and exceeded the value in every sampled ED subset. However, the corresponding PCA comparison did not support an ET advantage in concentration, so greater pairwise alignment did not establish that ET gradients occupied a lower-dimensional subspace. These results concern the initial diagnostic, which is separate from the paired normalization intervention in Section 4; Figure 2 reports only the paired intervention values.

B.2 Normalization Interventions

For the paired intervention, the same pre-update parameters and input batch are used in both computations. BATCH uses current batch means and variances, whereas FROZEN uses pretrained running means and variances. Source-defined ET/ED membership remains fixed. Only the BATCH gradient is used to advance the canonical trajectory; FROZEN does not generate an independently evolving adaptation trajectory.

A separate forward-only control evaluated 54 batches at unchanged source weights. Using pretrained running statistics recovered the source predictions on all 864 sample appearances. This control concerns prediction consistency under different normalization computations. It does not measure gradient coherence or update utility.

The factorial intervention separates the sources of normalization means and variances across the four conditions in Table 7, with BB and SS corresponding to BATCH and FROZEN, respectively. All four computations use the current canonical affine parameters, and any required batch statistics are computed from the activations produced by that condition. In particular, SB and BS do not reuse statistics cached from the BB forward pass, because changing normalization at an earlier layer changes the activations reaching later layers. The intervention therefore operates across the network rather than independently replacing statistics at one isolated layer.

Table 7: Sources of normalization statistics in the factorial intervention.

Condition	Mean statistics	Variance statistics	Advances trajectory
BB	Current batch	Current batch	Yes
SB	Pretrained running	Current batch	No
BS	Current batch	Pretrained running	No
SS	Pretrained running	Pretrained running	No

B.3 Directional Structure and Factorial Decomposition

The paired gradient comparison shows substantial reorientation alongside a change in norm (Table 8). These values summarize paired observations from the same canonical trajectory. The norm ratio is computed separately for each pair before aggregation.

Table 8: Direction and magnitude changes between paired gradient computations.

Paired quantity	Mean	Median
BATCH–FROZEN cosine	0.010982	0.010098
FROZEN-to-BATCH norm ratio	2.524677	2.469931

To examine shared directional structure, the 247 ED observations are divided into five folds using either random assignment or contiguous temporal blocks. For each held-out fold, the mean FROZEN gradient is estimated using only the other ED folds. Two transformations are then applied to held-out gradients. Fixed-vector subtraction removes this training mean directly. Axis removal instead removes each gradient's projection onto the direction of that mean, allowing the removed component to vary in magnitude across observations.

Table 9 reports means of the held-out, within-fold pairwise cosine summaries. Both transformations remove substantial shared structure under random folds. Temporal blocking reveals a larger difference: in the earliest held-out block, coherence decreases by 0.0303 after fixed-vector subtraction and by 0.2956 after axis removal. A shared direction therefore generalizes beyond the observations used to estimate it, while its strength varies along the trajectory. This result does not assign utility to that direction.

Table 9: Held-out coherence after removing an estimated shared vector or axis.

Fold construction	Raw FROZEN	Fixed-vector subtraction	Axis removal
Random assignment	0.5612	0.0270	0.0229
Temporal blocks	0.6023	0.2343	0.1504

For the factorial analysis, let $\mathbf{g}_n^b$ denote observation n computed under condition b at the same pre-update state and batch. The full change and its components are defined as

$$\begin{aligned} \boldsymbol{\Delta}_{\mathrm{full},n} &= \mathbf{g}_n^{\mathrm{SS}} - \mathbf{g}_n^{\mathrm{BB}}, \\ \boldsymbol{\Delta}_{\mathrm{mean},n} &= \mathbf{g}_n^{\mathrm{SB}} - \mathbf{g}_n^{\mathrm{BB}}, \\ \boldsymbol{\Delta}_{\mathrm{var},n} &= \mathbf{g}_n^{\mathrm{BS}} - \mathbf{g}_n^{\mathrm{BB}}, \\ \boldsymbol{\Delta}_{\mathrm{int},n} &= \mathbf{g}_n^{\mathrm{SS}} - \mathbf{g}_n^{\mathrm{SB}} - \mathbf{g}_n^{\mathrm{BS}} + \mathbf{g}_n^{\mathrm{BB}}. \end{aligned} \tag{12}$$

These definitions give the exact identity

$$\boldsymbol{\Delta}_{\mathrm{full},n} = \boldsymbol{\Delta}_{\mathrm{mean},n} + \boldsymbol{\Delta}_{\mathrm{var},n} + \boldsymbol{\Delta}_{\mathrm{int},n}. \tag{13}$$

The interaction term captures the part of the joint change that is not reconstructed by adding the two individual interventions. It is a network-level contrast, not a product of local mean and variance effects.

For each held-out observation, each component's signed projection onto the training-derived SS mean direction is divided by the full change's signed projection onto that direction. The resulting ratios describe contributions along the shared direction; they are not fractions of total gradient energy. Their means across the five temporal test folds are reported in Table 10.

Table 10: Mean signed contributions along the shared direction in temporal test folds.

Temporal fold	Mean component	Variance component	Interaction component
1	0.427	0.122	0.451
2	0.285	0.075	0.640
3	0.216	0.055	0.729
4	0.171	0.041	0.787
5	0.155	0.028	0.817

Figure 5 summarizes the held-out means and signed factorial contributions, showing that the interaction contribution increases across the temporal folds while the mean component is more prominent early in the trajectory. Neither individual intervention accounts for the full transformation. These results do not identify a unique responsible layer or establish a complete explanation of gradient coherence.

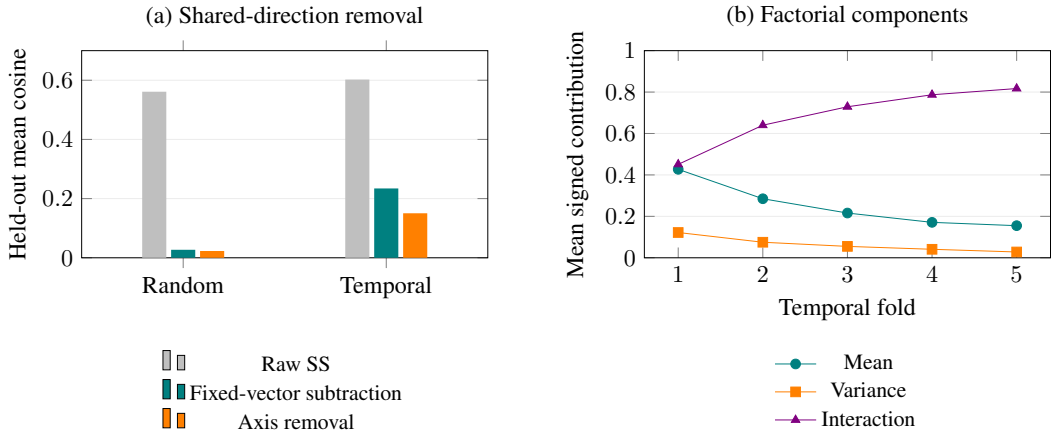

Figure 5: Shared-direction removal reduces held-out coherence; factorial contributions vary across temporal folds.

B.4 Additional Corruption Conditions

The paired BB–SS comparison is repeated under four additional ImageNet-C corruptions at severity 5. Each run starts from the source model and processes 5,000 images with batch size 16. The primary analysis uses the first 256 gradient positions.

Only batches assigned entirely to ET or ED enter the group comparison. Each corruption has one mixed batch in this window, which is excluded. ED subsets are matched to the corresponding ET group size, and the same subset identities are used for the paired BB and SS comparison.

Each gap in Table 11 is ET coherence minus the mean coherence across 1,000 matched ED subsets, and gap change is SS minus BB. The gap decreases under shot noise but increases under the other three corruptions. Across matched draws, the change is negative in 93.0% of shot-noise comparisons and positive in 100%, 100%, and 91.6% of motion-blur, snow, and contrast comparisons, respectively. These percentages describe overlapping subset comparisons, not independent success rates. The extension thus supports recurring sensitivity to gradient computation without establishing a uniform direction of change or a reversal under every corruption.

Table 11: Coherence gaps under additional corruptions. Gap change is SS minus BB.

Corruption	ET / ED gradients	BB gap	SS gap	Gap change
Shot noise	11 / 244	$+0.0311$	-0.0355	-0.0666
Motion blur	45 / 210	$+0.0284$	$+0.1310$	$+0.1025$
Snow	51 / 204	$+0.0280$	$+0.0809$	$+0.0529$
Contrast	16 / 239	$+0.0323$	$+0.0865$	$+0.0542$

C Exploratory and Observational Diagnostics

These supporting diagnostics examine candidate explanations for shared gradient structure and its variation along an adaptation trajectory. Their observational and exploratory scope is retained in the interpretation of each result.

C.1 Shared Source Predictions

An exploratory comparison examined whether ED images sharing the same incorrect source prediction produced more coherent gradients than ED images with diverse source predictions. The design used nine focal-class replicates, with six batch gradients per condition and replicate. Source parameters were fixed, and no optimizer updates were applied.

The mean paired coherence difference was -0.004154, with positive differences in three of nine replicates. The bootstrap interval included zero. However, subsequent checks showed that the shared source prediction was generally not preserved during the batch-dependent forward pass used to compute gradients. Most batches constructed to share a source prediction instead produced 16 distinct predicted labels during that pass.

A follow-up search attempted to construct batches that preserved the shared prediction under the gradient-generating computation, but the bounded search did not find enough eligible batches for the planned gradient comparison. This difficulty, together with the failure to maintain the intended shared-prediction condition in the initial comparison, motivated the normalization controls in Appendix B. These observations do not establish that systematic prediction errors cannot produce coherent gradients.

C.2 Affine Drift and Normalization Mismatch

Two observational screens examined whether simple trajectory-level measurements tracked the interaction contribution defined in Appendix B.3. Both used the same 247 ED positions. Table 12 defines the measurements.

Table 12: Measurements used in the observational screens.

Screen	Measurement
Affine drift	Separate root-mean-square (RMS) deviations of the normalization scale γ and shift β from their source values
Mean mismatch	Batch-minus-source activation means, normalized by source standard deviations
Variance mismatch	Log ratios of batch to source activation variances

Affine drift is measured as an absolute parameter deviation, without division by source parameter magnitude. Normalization mismatch is measured at the inputs to BatchNorm layers during the native BB computation. Channel-level measurements are combined using the specified hierarchical RMS aggregation.

For each screen, associations with the interaction contribution are examined before and after removing a quadratic trend in stream position from both quantities. Spearman correlations are then

calculated between the residuals (Table 13). This detrending is descriptive and does not constitute causal adjustment.

Table 13: Residual Spearman correlations after removing quadratic trends in stream position.

Measurement	Random folds	Temporal folds
Scale-parameter drift	-0.196	-0.203
Shift-parameter drift	-0.187	-0.193
Global mean mismatch	-0.077	-0.073
Global variance mismatch	-0.072	-0.078

Affine drift showed strong raw positive associations, approximately 0.78, but the positive direction did not persist after detrending, while global normalization mismatch showed only weak residual associations. These patterns do not provide a clear global explanation beyond coarse temporal progression. Because the observations come from a single evolving trajectory, they are not treated as independent replications, and the screens do not exclude a causal role for affine drift or normalization mismatch.

D Structural Retention and Cross-State Utility

This appendix examines whether low-dimensional gradient structure distinguishes source-defined sample groups and whether it is associated with the utility of native adaptation updates. These supporting analyses compare observations collected along an evolving trajectory under the settings in Table 14. They characterize structural discrimination and cross-state association, complementing the within-state directional ranking evaluated in Section 6.

D.1 Cross-fitted Structural Retention

The first 256 gradient observations were divided into five contiguous temporal folds. For each held-out fold f, a rank-25 principal component basis $\mathbf{Q}^b_{-f}$ was fitted using the other four folds under computation condition b. The corresponding training-gradient mean is denoted by $\overline{\mathbf{g}}^b_{-f}$. For observation n in the held-out fold, with a finite, nonzero gradient after centering by the training mean, cross-fitted retention is defined as

$$R^{\mathrm{cf},b}_n = \frac{\left\| (\mathbf{Q}^b_{-f})^\top (\mathbf{g}^b_n - \overline{\mathbf{g}}^b_{-f}) \right\|_2}{\left\| \mathbf{g}^b_n - \overline{\mathbf{g}}^b_{-f} \right\|_2}. \tag{14}$$

This score measures the fraction of the centered gradient norm retained by the fitted subspace. It differs from C_j, which measures squared retention of an uncentered candidate gradient under a projector estimated from strictly preceding history. Cross-fitting excludes the held-out fold from basis estimation, but its training folds may contain later observations. It is therefore a retrospective structural diagnostic rather than an online coherence score.

D.2 Retention Distinguishes Groups under Both Computations

Structural discrimination was evaluated using the difference between the median ET retention and the median retention of each size-matched ED subset. The same subset identities were used under BB and SS, allowing the change in this difference to be evaluated within each matched draw (Table 15). The ET median retention was 0.1076 under BB and 0.8364 under SS. The mean of the matched ED medians was 0.1026 and 0.7882, respectively. Thus, positive ET–ED discrimination was retained under both computations, while the difference was larger under SS in most matched draws.

Figure 6 shows the empirical distributions over matched draws and the fold-wise associations described below. The group comparisons concern structural discrimination rather than update quality, and no principled effect-size threshold was available to establish whether the change was materially

Table 14: Configuration of the structural-retention and cross-state utility analyses.

Component	Setting
Model and corruption	ResNet-50; Gaussian noise, severity 5
Gradient observations	First 256 adaptation positions
Cross-fitting	Five contiguous temporal folds
Subspace	Rank-25 PCA, fitted separately for each held-out fold and computation condition
Centering	Mean gradient from the four training folds
Computation conditions	BB: current-batch normalization statistics; SS: pretrained running statistics
Diagnostic groups	Source-defined entropy-truthful (ET) and entropy-deceptive (ED) groups
Group comparison	1,000 ED subsets matched to the ET group size
Utility evaluation	Native BB updates; a fixed probe of 904 images outside the 4,096 adaptation images
Probe batching	57 fixed batches under current-batch normalization

Table 15: Retention differences across matched ED subsets under the two computation conditions.

Comparison	Mean	Median	Positive draws
ET minus matched ED, BB	0.0050	0.0054	81.6%
ET minus matched ED, SS	0.0482	0.0434	87.4%
Change in the ET–ED difference, SS minus BB	0.0432	0.0400	82.8%

important. Because matched ED subsets overlap, their positive fractions summarize subset comparisons rather than independent replication rates or significance estimates. The displayed distributions are descriptive, not confidence intervals, and norm-controlled points are offset horizontally for visibility.

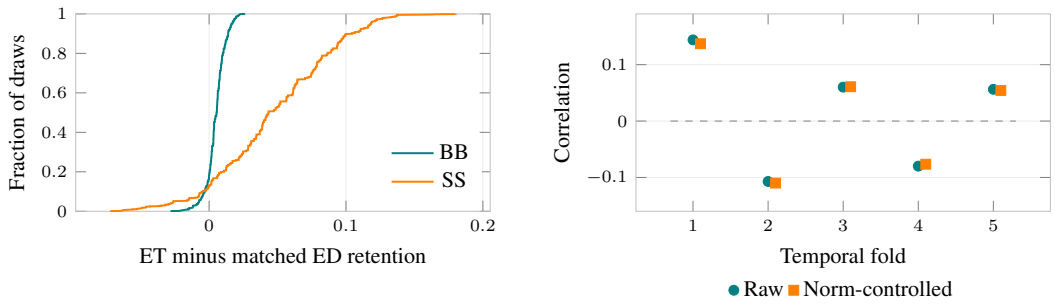

Figure 6: Structural discrimination persists, while retention–utility associations vary in sign across temporal folds.

D.3 Association with Native One-step Utility

The BB retention scores were subsequently compared with the supervised utility of the corresponding native adaptation updates. At each position, utility was measured as the mean probe cross-entropy before the update minus that after the update. Positive values therefore indicate improvement. Probe identities, order, and batch partition were fixed throughout evaluation.

Unlike the directional comparison in Section 6, these observations arise from different model states, and their update norms are not equalized. The analysis asks whether retention and native update utility vary together along the trajectory, rather than whether retention ranks alternative directions at a shared state.

Spearman correlations were computed separately within each temporal fold (Table 16). A gradient-norm-controlled association was also calculated by regressing the retention ranks and utility ranks separately on the ranked log gradient norm, including an intercept, and correlating the two residual vectors. Both median associations were positive, but each analysis had positive correlations in only

Table 16: Fold-wise associations between cross-fitted retention and native one-step supervised utility.

Temporal fold	Adaptation positions	Spearman	Norm-controlled
1	1–52	0.144	0.137
2	53–103	−0.107	−0.110
3	104–154	0.060	0.061
4	155–205	−0.080	−0.077
5	206–256	0.056	0.054
Median across folds		0.056	0.054
Folds with positive association		3/5	3/5

three of five folds. The predefined criterion required a positive median and at least four positive folds for both analyses. A temporally stable positive association was therefore not established.

This result does not show that retention and utility are independent or that gradient structure cannot support useful adaptation. Because the comparison spans changing model states and native updates, it cannot determine whether a more coherent direction is preferable at a given state. That unresolved distinction motivates the controlled same-state comparison in Section 6.

E Projector and Trajectory Controls

E.1 Intervention Construction

The projector comparison is conducted along one canonical adaptation trajectory (Table 17). At each historical sampling position, gradients are computed using paired normalization conditions at the same model state and on the same batch. BB histories use current batch statistics. SS histories use pretrained running statistics.

The full SS computation applies its own sample selection and entropy weights. In the selection-fixed computation, the exact sample identities selected by BB and their detached BB weights are retained. Thus, the two SS interventions distinguish the combined effect of computation and selection from the effect that remains when selection and weighting are fixed.

Table 17: Configuration of the frozen projector comparison.

Setting	Specification
Model and corruption	Standard ResNet-50; Gaussian Noise, severity 5
Stream	50,000 unique images; label-free permutation, seed 1
Batch size	16
Adapted parameters	BatchNorm affine parameters
Subspace rank	25
History sampling interval	Every 50 batches
Maximum history size	100 gradient snapshots
Recorded historical snapshots	62
Frozen projector comparisons	20 temporal anchors

For each history, PCA is fitted after centering its stored gradients. The resulting orthogonal projector is applied to the same saved raw BB gradient without subtracting the historical mean from that gradient. The projectors are denoted by $\mathbf{P}_a^{\mathrm{BB}}$, $\mathbf{P}_a^{\mathrm{SS,full}}$, and $\mathbf{P}_a^{\mathrm{SS,fixed}}$.

At an anchor, the current gradient is included in the history under the native tracking convention, unlike the strictly-pre-anchor history used in Section 6. The present comparison therefore measures the operational effect of replacing the projector rather than held-out subspace generalization. Since the anchors share histories along one trajectory, their variation is reported descriptively without treating them as independent replications.

E.2 Metrics and Controls

Section 5 defines directional discrepancy D_{proj}. To assess the magnitude of the change, the projected-gradient difference is also normalized by the raw-gradient norm. For the selection-fixed intervention,

$$R^{\mathrm{fixed}}_{\mathrm{proj},a} = \frac{\left\|\mathbf{P}_a^{\mathrm{BB}}\mathbf{g}_a^{\mathrm{BB}} - \mathbf{P}_a^{\mathrm{SS,fixed}}\mathbf{g}_a^{\mathrm{BB}}\right\|_2}{\left\|\mathbf{g}_a^{\mathrm{BB}}\right\|_2}. \tag{15}$$

The same definition is used for the full, replay, and random comparisons with the corresponding alternative projector.

BB replay repeats the BB history construction and projection. The random control replaces the alternative basis with a rank-25 orthonormal basis obtained from the orthonormal factor of a QR decomposition of a Gaussian random matrix. It provides context for directional discrepancies in the adapted-parameter space, rather than a utility benchmark.

Table 18 reports medians and interquartile ranges across the 20 anchors, with Q1 and Q3 denoting the first and third quartiles. Both SS interventions exceed the paired replay values for both metrics at every anchor. The maximum replay directional discrepancy is 1.11×10^{-16}, while its relative difference is zero throughout.

Table 18: Projected-gradient discrepancies across 20 anchors: median and interquartile range.

Comparison	D_{proj}, median [Q1, Q3]	R_{proj}, median [Q1, Q3]
Full	0.9484 [0.9308, 0.9734]	0.4195 [0.1635, 0.9820]
Selection-fixed	0.9547 [0.9322, 0.9772]	0.4194 [0.1640, 0.9831]
BB replay	≈ 0 [0, 1.11×10^{-16}]	0 [0, 0]
Matched-rank random	0.9900 [0.9805, 0.9945]	0.4170 [0.1434, 0.9848]

Figure 3(a) displays every anchor, with vertical offsets used only to separate overlapping points. Black marks indicate medians. The full and selection-fixed results are listed in Table 19; anchor positions identify stream locations rather than independent runs.

The native BB projection retains a median of 0.4167 of the raw-gradient norm, compared with 0.0707 for the full SS projection and 0.0715 for the selection-fixed SS projection. The directional result therefore accompanies a measurable magnitude difference, although native retained norm varies substantially across anchors, from 0.0867 to 0.9938, consistent with the variation in R_{proj}. Together, these measurements establish projector-dependent gradients without identifying either projector as more useful for the task.

E.3 Trajectory Comparisons

The trajectory experiment begins from complete states saved at three positions along one canonical BB adaptation stream (Table 20). At each start, the state is cloned into two branches with equal initial parameters, optimizer state, histories, and other adaptation state.

Arm A continues with its own BB-derived projector. Arm B uses an SS-derived projector whose historical gradients retain the samples and weights selected by BB at that branch's own states. Both branches compute their adaptation forward passes and raw gradients using BB normalization. After intervention, their states, gradients, and histories are allowed to diverge.

Complete-state equality and absence of shared mutable storage are checked before the branches evolve. The BB-to-BB controls produce zero prediction and accuracy differences, with zero recorded discrepancies in their paired parameters, gradients, histories, and optimizer momentum.

Table 19: Full and selection-fixed projected-gradient discrepancies at each anchor.

Anchor position	$D^{\text{full}}_{\text{proj}}$	$D^{\text{fixed}}_{\text{proj}}$	$R^{\text{full}}_{\text{proj}}$	$R^{\text{fixed}}_{\text{proj}}$
1300	0.9263	0.9236	0.9833	0.9831
1350	0.9423	0.9383	0.9750	0.9748
1450	0.9317	0.9327	0.9872	0.9873
1550	0.9742	0.9713	0.2280	0.2179
1650	0.8982	0.9157	0.9765	0.9781
1750	0.9928	0.9981	0.1412	0.1422
1850	0.9210	0.9432	0.9816	0.9831
1950	0.9627	0.9544	0.3168	0.3177
2050	0.9479	0.9549	0.4525	0.4522
2150	0.9489	0.9559	0.1347	0.1384
2200	0.9211	0.9476	0.9794	0.9811
2300	0.9283	0.9213	0.9912	0.9907
2400	0.9333	0.9308	0.9871	0.9870
2500	1.0112	0.9921	0.1677	0.1693
2600	0.9846	0.9967	0.1512	0.1484
2700	0.9655	0.9755	0.3864	0.3867
2800	0.9642	1.0004	0.1813	0.1757
2900	0.9893	0.9824	0.1211	0.1187
3000	0.9365	0.9278	0.9899	0.9893
3100	0.9731	0.9598	0.1307	0.1302

Table 20: Configuration and endpoints of the trajectory comparison.

Setting	Specification
Stream and model	ResNet-50; 50,000 Gaussian Noise images; batch size 16
Intervention starts	Zero-based batch indices 1300, 1900, 2500
Final batch index	3124
Arm A	Own-state BB gradients and BB history
Arm B	Own-state BB gradients and selection-fixed SS history
Prediction evaluation	Online top-1 predictions before each update
Primary endpoint	Mean accuracy over all suffix images
Materiality threshold	Absolute accuracy difference of 1 percentage point
Negative control	BB-to-BB replay over each complete suffix

Table 21 reports suffix-level accuracy and prediction disagreement, both calculated over all images within each suffix. Disagreement counts images for which the two top-1 predictions differ. The three suffixes overlap and are reported separately.

Table 21: Online accuracy and prediction disagreement over each intervention suffix.

Start	Suffix images	Accuracy A (%)	Accuracy B (%)	B−A (pp)	Disagreement (%)
1300	29,200	30.0514	30.2295	+0.1781	10.4281
1900	19,600	30.3776	30.4133	+0.0357	8.8520
2500	10,000	30.6600	30.7600	+0.1000	8.4500

Parameter divergence is measured after each update as

$$\left\|\boldsymbol{\theta}_t^{\mathrm{A}} - \boldsymbol{\theta}_t^{\mathrm{B}}\right\|_2, \tag{16}$$

The branch vectors follow the notation introduced in Section 5.2. This unnormalized Euclidean distance is computed over the adapted normalization-affine parameters.

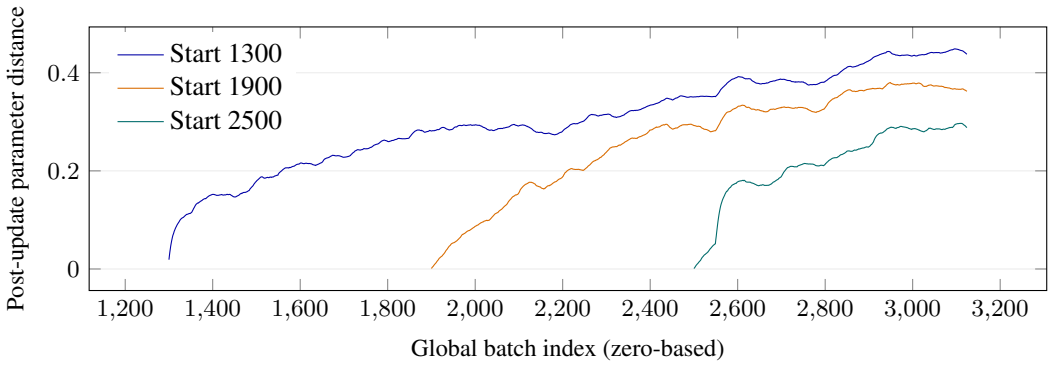

Figure 7: Adapted parameter trajectories diverge after projector interventions from three shared starting states.

The curves in Figure 7 show recorded per-step distances without smoothing. Together with prediction disagreement, they demonstrate that the projector intervention affects the evolving adaptation process. The small accuracy differences do not imply that the branches remain functionally identical.

E.4 Prediction Decomposition

Each suffix image is assigned to one of five mutually exclusive categories according to the two predictions and the true label (Table 22). Disagreements comprise the two correctness-transition categories and the category in which both predictions are wrong but different. This last category accounts for 75.9%, 76.4%, and 77.5% of disagreements at the three starts, respectively.

Table 22: Prediction categories and disagreement counts for each intervention suffix.

Prediction category	Start 1300	Start 1900	Start 2500
Both correct	8,434	5,753	2,976
A correct, B wrong	341	201	90
A wrong, B correct	393	208	100
Both wrong, same prediction	17,721	12,112	6,179
Both wrong, different predictions	2,311	1,326	655
Total images	29,200	19,600	10,000
Total disagreements	3,045	1,735	845

Figure 3(b) expresses each disagreement category as a percentage of all suffix images, so the stacked bar length equals the overall disagreement rate. The percentages above instead use the number of disagreements as their denominator.

Relative to Arm A, Arm B gains 393 correct predictions and loses 341 at start 1300, yielding a net gain of 52. The corresponding gains and losses are 208 versus 201 at start 1900 and 100 versus 90 at start 2500. Dividing these net gains by the respective suffix sizes gives the accuracy differences in Table 21.

This decomposition explains how substantial prediction disagreement can coexist with small aggregate accuracy differences. It does not imply that the disagreements are unimportant or that coherence is uninformative. Rather, behavioral change and task improvement must be evaluated separately.

Figure 8 preserves anchor order for both metrics and all controls. Lines connect observations along the shared trajectory; they do not indicate independent replications.

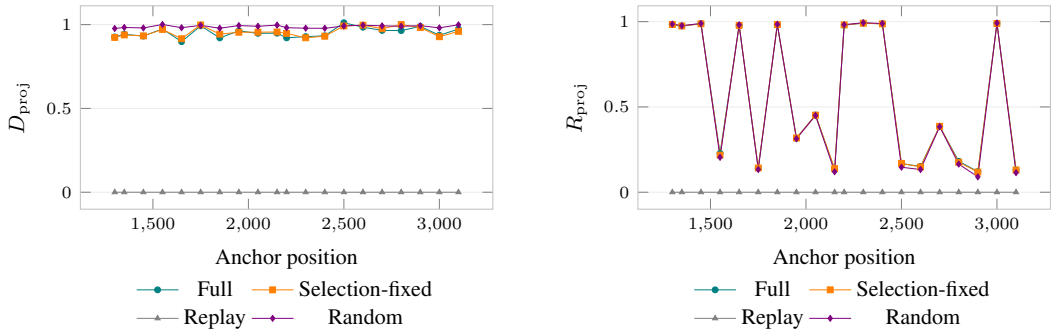

Figure 8: Directional discrepancies remain large, while relative differences vary substantially across the same temporal anchors.

F Directional Evaluation and Ranking Details

F.1 State Restoration and Perturbations

Each directional comparison starts from a complete canonical state saved immediately before an anchor batch. Restoration includes model parameters and buffers, optimizer momentum, gradient histories, probability-tracking state, stream position, random-number-generator state, and model execution modes. Candidate computations and probe evaluations do not persist changes to the canonical trajectory.

The three anchor positions are 1350, 1900, and 2500. For ViT, these positions are measured from the start of the target continuation. The index $a \in \{1, 2, 3\}$ identifies these anchors in the ranking equations.

The candidate and probe blocks in Table 23 are constructed before evaluating utility and do not overlap within an anchor. The specified blocks are also disjoint across the three anchors of a trajectory. This construction defines an offline comparison on future local data rather than an online direction-selection procedure.

Table 23: Candidate and probe construction relative to each anchor.

Component	Construction relative to the anchor batch
Reference history	Tracked gradient snapshots strictly before the anchor
Candidates	32 batches at offsets 0–31
Evaluation probe	64 batches at offsets 32–95
Candidate batch size	16
Probe size	1,024 images
Candidate state	Independently restored reference state
Perturbation scale	Norm of the native anchor-update displacement

The projector is obtained from centered historical gradients. In the ResNet-50 regime, its rank is 25 and at least 26 preceding snapshots are required. In the ViT regime, the rank is 50 and at least 51 preceding snapshots are required. Rank and projector checks are performed before candidate evaluation; a deficient history is not repaired by reducing the requested rank.

Candidate gradients are computed from the native adaptation objective before projection. Coherence is evaluated using Eq. (1), and each original gradient direction is applied through the equal-norm perturbation in Eq. (6). These operations separately measure compatibility with the historical subspace and the effect of following the candidate direction.

To obtain δ_{ra}, the canonical anchor batch is processed on an isolated state copy using the native tracking, projection, and optimizer sequence. Only the resulting displacement norm is retained for

the directional comparison. Each candidate is then applied directly as specified in Section 6, without a candidate-specific optimizer or momentum step.

Baseline and perturbed predictions use identical probe batches. ResNet-50 evaluation retains current-batch normalization, making the fixed batch partition part of the evaluation definition. Probe-induced state changes are discarded. Cross-entropy and forward KL are averaged over images, with the same weighting for baseline and candidate predictions.

F.2 Ranking, Ties, and Aggregation

The eligible set for utility concordance is

$$\mathcal{J}_{ra}^{CU} = \{(j,k) : j < k,\ |C_{raj} - C_{rak}| > \varepsilon_C,\ |U_{raj} - U_{rak}| > \varepsilon_U\}. \tag{17}$$

The influence set $\mathcal{J}_{ra}^{CI}$ replaces the utility condition with

$$|I_{raj} - I_{rak}| > \varepsilon_I. \tag{18}$$

Thus, the two comparisons use their own eligible sets rather than assuming identical denominators.

Tie tolerances (Table 24) are fixed before the full runs. The coherence tolerance is eight times single-precision machine epsilon. Influence and utility tolerances are the larger of eight times double-precision machine epsilon and twice the maximum discrepancy observed in deterministic replay.

The replay panel uses seed 1, all three anchors, and candidate offsets 0, 15, and 31. Each case is evaluated twice from the identical full state with the identical direction and probe. For ViT, calibration uses the Gaussian-noise setting and is fixed before evaluating both target conditions.

Table 24: Numerical tie tolerances, rounded for presentation.

Tolerance	ResNet-50	ViT
ε_C	9.54×10^{-7}	Same
ε_I	1.78×10^{-15}	Same
ε_U	1.78×10^{-15}	Same

Repeated influence and utility evaluations had zero recorded discrepancies in the calibration panels, so their numerical floors determined the tolerances. These tolerances address numerical ties; they are not thresholds of practical importance.

With 32 candidates, each anchor has 496 possible unordered pairs. One coherence-tied pair is excluded at the ResNet-50 anchor at position 2500 in trajectory 4, leaving 495 eligible pairs for both comparisons. All other ResNet-50 anchors and all ViT anchors retain 496 pairs. No influence or utility ties are excluded in these recorded results.

Aggregation preserves the experimental hierarchy:

$$q_r^{CU} = \frac{1}{3}\sum_{a=1}^{3} q_{ra}^{CU}, \qquad \bar{q}^{CU} = \frac{1}{5}\sum_{r=1}^{5} q_r^{CU}. \tag{19}$$

The same hierarchy is used for C–I and the reported secondary summaries. The two ViT targets are summarized separately. Within a target, the continuation is the highest-level varying unit, conditional on its shared warm-up.

Candidate evaluation requires finite, nonzero gradients in both regimes. The ViT protocol additionally applies its calibrated near-zero threshold and does not replace invalid candidates with later batches. Required concordances with no eligible pairs invalidate the corresponding evaluation rather than being omitted from aggregation.

At each anchor, the transformed concordance is $\tau_{ra}^{CU} = 2q_{ra}^{CU} - 1$, with the corresponding definition for C–I. These transformed scores and within-anchor Spearman correlations are secondary summaries, aggregated using the same anchor-then-run hierarchy. No primary significance test or equivalence test is attached to the reported concordances.

F.3 RESNET-50 RESULTS

Five label-free permutations produce 15 frozen states and 480 candidate directions. The trajectory means in Table 25 average the three anchors equally.

Table 25: ResNet-50 trajectory-level ranking summaries.

Trajectory	C–I concordance	C–U concordance	Spearman C–I	Spearman C–U
1	0.7063	0.4200	0.5663	-0.2600
2	0.6922	0.4940	0.5329	-0.0365
3	0.7063	0.5034	0.5530	-0.0011
4	0.6429	0.4499	0.3939	-0.1399
5	0.6942	0.5087	0.5382	0.0255
Equal-weight mean	0.6884	0.4752	0.5169	-0.0824

All five trajectory-level C–I means exceed 0.5. The C–U means fall on both sides of 0.5, consistent with the absence of a uniformly positive utility-ranking pattern. Spearman summaries show the same broad contrast, while measuring rank association rather than pairwise concordance under the calibrated eligibility rules.

For completeness, the per-anchor concordances for all three settings are reported together in Table 26. GN denotes Gaussian noise and MB denotes motion blur. The seed identifies a ResNet-50 trajectory or a ViT continuation within its target.

Table 26: Concordances at all 45 anchors, grouped by setting.

		R50 GN		ViT GN		ViT MB	
Seed	Anchor	C–I	C–U	C–I	C–U	C–I	C–U
1	1350	0.724	0.419	0.815	0.486	0.692	0.528
1	1900	0.742	0.385	0.794	0.528	0.829	0.530
1	2500	0.653	0.456	0.875	0.488	0.851	0.478
2	1350	0.657	0.506	0.802	0.528	0.954	0.675
2	1900	0.726	0.440	0.802	0.569	0.790	0.544
2	2500	0.694	0.536	0.827	0.667	0.802	0.558
3	1350	0.714	0.536	0.839	0.554	0.764	0.486
3	1900	0.702	0.478	0.869	0.405	0.879	0.456
3	2500	0.704	0.496	0.829	0.591	0.821	0.512
4	1350	0.661	0.417	0.675	0.577	0.819	0.558
4	1900	0.589	0.544	0.724	0.558	0.835	0.514
4	2500	0.679	0.388	0.863	0.478	0.859	0.472
5	1350	0.649	0.458	0.810	0.371	0.808	0.391
5	1900	0.712	0.571	0.831	0.446	0.770	0.554
5	2500	0.722	0.498	0.857	0.484	0.750	0.450

Figure 4 connects the C–I and C–U scores from the same anchor. Small horizontal offsets separate overlapping observations and do not encode another quantity. Colors identify trajectories or continuations within each setting, not paired observations across different models. Black bars show the hierarchical aggregate means, not confidence intervals.

F.4 VIT PROTOCOL AND RESULTS

Each target condition begins independently from the pretrained source model (Table 27). Warm-up follows the fixed corruption order with the target removed. After completing the 14 remaining corruption segments, the sequence restarts and stops after 17 segments in total.

Each segment contains 5,000 sample appearances at severity 5. With batch size 16 and a final partial batch per segment, warm-up comprises 85,000 sample appearances and 5,321 batches, with no model reset between segments. Within a target, all five continuations share this deterministic warm-up state and differ in their target-data permutation. Appendix A.3 documents this batch-size choice and its difference from the original ViT benchmark.

Table 27: ViT configuration for the directional comparison.

Setting	ViT specification
Model	ViT-B/16
Adapted parameters	LayerNorm-affine, 38,400 scalars
Target conditions	Gaussian noise; motion blur, severity 5
Target stream per continuation	50,000 images
Batch size	16
Subspace rank / history limit	50 / 100
History sampling interval	Every 100 batches
Learning rate	5×10^{-4}
Target-local anchor positions	1350, 1900, 2500
Continuations	Five per target, seeds 1–5

The continuation-level concordances in Table 28 are complemented by mean within-anchor Spearman correlations of 0.7979 for C–I and 0.0424 for C–U under Gaussian noise, and 0.7993 and 0.0337 under motion blur. These summaries retain the contrast between influence and utility ranking without reproducing the negative aggregate C–U sign observed for ResNet-50.

Table 28: ViT continuation-level concordances for each target.

Continuation	GN C–I	GN C–U	MB C–I	MB C–U
1	0.8280	0.5007	0.7903	0.5121
2	0.8105	0.5880	0.8488	0.5927
3	0.8454	0.5168	0.8212	0.4845
4	0.7540	0.5376	0.8374	0.5148
5	0.8327	0.4335	0.7762	0.4651
Equal-weight mean	0.8141	0.5153	0.8148	0.5138

The continuations assess variation in target order conditional on a shared warm-up. They do not measure variation across independently generated warm-up histories. Furthermore, architecture, normalization, learning rate, subspace rank, sampling cadence, and preceding history differ from the ResNet-50 regime. The observed quantitative differences cannot be attributed to any one of these factors.