MORE DATA, WORSE DECISIONS? EXPLAINING PREFERENCE REVERSALS IN NEURAL NETWORKS THROUGH REPRESENTATION GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks increasingly incorporate additional information from multiple sources to improve generalization, but combining such information can also alter previously consistent decisions. While more data provides richer information, information integration may disrupt previously consistent decisions and induce decision instability. This raises the question of why additional information can sometimes lead to worse decisions by reversing previously consistent preferences. We study this question from the perspective of representation geometry. We first reveal that such preference reversals arise from changes in representation geometry during information integration. Building on this mechanism, we develop a geometric criterion for assessing whether candidate information sources can preserve decision stability before information integration. Finally, we show that representation geometry can be controlled during training to improve compositional stability under information integration. Experiments on controlled settings and real-world decision tasks support the identified mechanism, examine the relationship between geometric compatibility and decision stability, and evaluate how geometry-based interventions influence decision stability. These findings provide a geometric perspective for understanding and improving decision stability in neural network systems that integrate information from multiple sources.

1 INTRODUCTION

Neural networks are increasingly trained with information collected from multiple sources, allowing models to learn from more diverse information than any single source can provide. This paradigm has become common in areas such as multimodal learning, domain adaptation, and large-scale data integration, where combining complementary information often improves model performance (Zhao et al., 2024; Luo et al., 2024; Wen et al., 2024). However, integrating information from different sources does not only expand the information available to a model. It can also change how the model uses learned representations to support decision preferences.

Integrating information from different sources can therefore create a critical risk: the combined model may not preserve decisions that were previously consistent across individual sources. When independently trained models favor the same choice but the integrated model later selects a different one for the same decision problem, this change represents a preference reversal and leads to decision instability. Recent studies on multi-source adaptation have shown that differences between sources can introduce negative transfer and degrade predictive performance when information is combined without proper adaptation (Zhao et al., 2018). However, these studies mainly focus on improving adaptation strategies or recovering predictive performance after integration. They do not explain why combining information can change decisions that were already consistent across individual sources.

The phenomenon above raises a fundamental question: how should a neural network maintain decision stability when information from different sources is integrated? Case-Based Decision Theory (CBDT), introduced by Gilboa and Schmeidler, provides a theoretical perspective by requiring that preferences supported by separate cases remain consistent when the cases are combined (Gilboa & Schmeidler, 1995; Gilboa et al., 2002). Building on this principle, recent work connects neural network decisions with their training cases through an exact case-based decomposition, providing a way to analyze

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

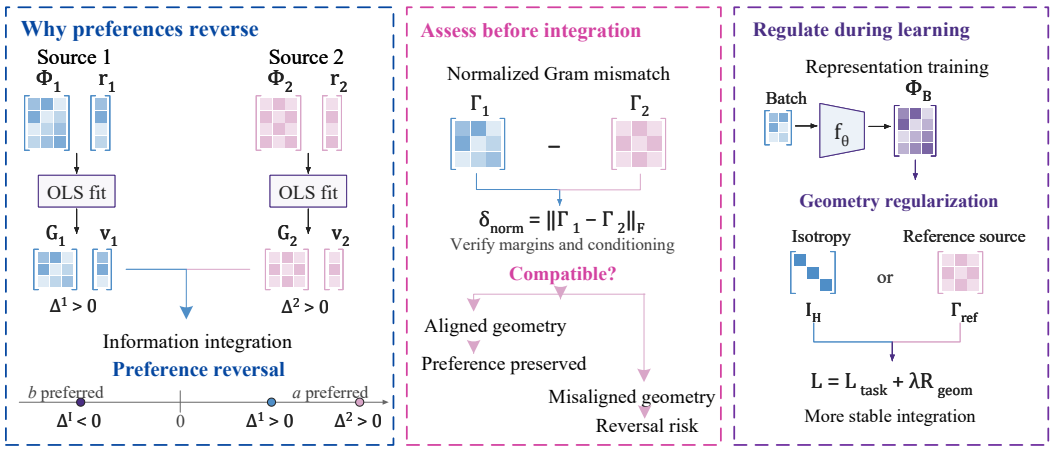


Figure 1: Overview of preference reversals under information integration and their analysis through representation geometry.

how learned representations support model decisions (Yan et al., 2026). Together, these studies provide a foundation for studying decision stability under information integration, but they do not explain how combining sources changes the representation geometry that supports neural network decisions. Understanding this process is essential for explaining why previously consistent decisions may become unstable after data integration.

In this paper, we study preference reversals in neural networks under information integration from the perspective of representation geometry. Our contributions are summarized as follows.

- First, we show how changes in representation geometry during information integration can reverse previously consistent decision preferences. This explains why preference reversals can arise even when individual models agree and how they can lead to decision instability.
- Building on this mechanism, we establish geometric conditions under which information integration preserves decision preferences. We further develop a geometry-based criterion for assessing integration stability before information integration.
- Finally, we show that representation geometry can be regulated during representation learning. This provides a principled way to enhance compositional stability under information integration.

Figure 1 summarizes the preference-reversal mechanism, the pre-integration assessment, and the geometry-regulation strategy.

2 Related Work

Existing studies have investigated neural network decisions from three complementary perspectives: how decisions are related to training information, how information is integrated across sources, and how representation structures organize learned information.

Training Cases and Decision Explanations. Research on neural network explanations has studied how learned models rely on training information to produce decisions. Influence-based approaches estimate the contribution of individual training examples to predictions (Koh & Liang, 2017; Pruthi et al., 2020), while concept- and example-based methods provide interpretable views of model decision mechanisms (Kim et al., 2018; Ghorbani et al., 2019; Yeh et al., 2018; Chen et al., 2019; Wolf et al., 2024). From a decision-theoretic perspective, CBDT explains decisions through supporting cases and requires preference consistency when cases are combined (Gilboa & Schmeidler, 1995; Gilboa et al., 2002). Recent work extends this perspective to neural networks by decomposing model outputs into their contributing training cases (Yan et al., 2026). These approaches have substantially improved our understanding of how individual models derive decisions from their

supporting information. Yet, decision processes across models trained with different information sources remain less understood.

Information Integration and Multi-Source Learning. Information integration methods aim to improve generalization by combining knowledge from heterogeneous sources. Multimodal learning investigates how complementary modalities can be combined, while domain adaptation and multi-source learning address distribution differences across sources (Baltrusaitis et al., 2019; Zhao et al., 2024; Ben-David et al., 2010; Ganin et al., 2016; Long et al., 2018). Subsequent studies introduce alignment and generalization strategies to reduce source discrepancy and improve robustness under distribution shifts (Sun et al., 2016; Zhao et al., 2018; Gulrajani & Lopez-Paz, 2021; He et al., 2023). Recent works further extend these approaches to more complex heterogeneous settings involving multiple sources and adaptation challenges (Li et al., 2024; Wen et al., 2024). Existing information integration methods have primarily been developed around improving prediction, generalization, and robustness across heterogeneous sources. However, how integration affects the resulting decisions of learned models beyond these performance-oriented objectives remains less understood.

Representation Geometry and Model Compatibility. Representation geometry provides a perspective for understanding how neural networks organize learned information. Previous studies analyze relationships between neural representations through geometric similarity measures, revealing shared structures across independently trained models (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019; Nguyen et al., 2021). Other studies show that controlling representation statistics can improve optimization stability and transferability through normalization and geometric alignment mechanisms (Ioffe & Szegedy, 2015; Huang et al., 2018; Sun et al., 2016). Recent model composition studies further demonstrate that representation compatibility influences whether independently learned models can be effectively combined (Matena & Raffel, 2022; Ainsworth et al., 2023; Ilharco et al., 2023; Yadav et al., 2023). Research on representation geometry has revealed important properties of learned representations through similarity, alignment, and compatibility measures. Despite these advances, how such geometric properties relate to the functional behavior of integrated models remains insufficiently characterized.

Together, these studies provide complementary views of decision formation, information integration, and representation organization. Establishing the relationship between representation changes during integration and the resulting decision outcomes remains an open challenge. In this work, we study this relationship from the perspective of representation geometry.

3 Preference Reversals under Information Integration

This section analyzes how information integration changes neural network decisions through representation geometry. It first formulates decision formation before and after integration, then explains how geometry changes can induce preference reversals, and finally distinguishes preference reversal from decision instability.

3.1 Decision Formation before and after Information Integration

This section describes how neural networks form decisions before and after information integration. For a neural network with a fixed representation function $\phi(\cdot)$, the decision rule evaluates candidate actions based on the learned representation. Here, fixed means that the same representation map is used when fitting each source and the combined cases. For input x, $\phi(x)$ denotes its neural representation, and $\hat{v}_a^M$ denotes the output parameter associated with action a in model M. The score assigned to action a is $Q^M(x,a) = \phi(x)^\top \hat{v}_a^M$, where $Q^M(x,a)$ represents how model M evaluates action a.

To compare two candidate actions a and b, we define the preference margin:

$$\Delta_{a,b}^M(x) = Q^M(x,a) - Q^M(x,b), \tag{1}$$

where $\Delta_{a,b}^M(x)$ measures the preference difference between actions a and b. A positive value indicates that model M prefers action a, while a negative value indicates a preference for action b.

The preference margin depends on the representations of training cases. Let Φ contain one case representation per row, and let $r_{a,b}$ contain the case-wise return of action a minus that of action b.

Under ordinary least squares (OLS) fitting with a full-column-rank Φ, the fitted coefficient difference $v^M = \hat{v}_a^M - \hat{v}_b^M$ equals $G^{-1}\Phi^\top r_{a,b}$. The matrix G is the Gram matrix of case representations:

$$G = \Phi^\top \Phi. \tag{2}$$

This formulation shows that decisions depend on both the case-wise returns and the geometry of their representations. Substituting the fitted coefficients gives $\Delta_{a,b}^M(x) = \phi(x)^\top G^{-1}\Phi^\top r_{a,b}$: the inverse Gram matrix transforms the return evidence before the query is scored.

When cases from multiple sources are integrated, the decision rule is refitted on their combined representations. Let Φ_I be the stacked case-representation matrix and $G_I = \Phi_I^\top \Phi_I$ its Gram matrix. The integrated preference margin is:

$$\Delta_{a,b}^I(x) = \phi(x)^\top G_I^{-1}\Phi_I^\top r_{a,b}, \tag{3}$$

where $\Delta_{a,b}^I(x)$ represents the preference margin after information integration. Compared with either source model, the integrated readout acts on the return evidence through a different inverse Gram matrix. This change provides a route from consistent source preferences to a different integrated preference. A complete notation table is provided in Appendix A.

3.2 Representation Geometry and Preference Reversals

Information integration can change preferences because the integrated model forms decisions under a different representation geometry. For the same action pair (a, b), source $j \in \{1, 2\}$ has case-representation matrix Φ_j, Gram matrix $G_j = \Phi_j^\top \Phi_j$, and return-difference vector $d_j(a, b)$ for its own cases. Its evidence vector is $y_j = \Phi_j^\top d_j(a, b)$, so its margin is $\Delta_{a,b}^j(x) = \phi(x)^\top G_j^{-1} y_j$. These margins describe how each model evaluates the same decision problem under its own representation geometry.

After integration, stacking the source cases gives $G_I = G_1 + G_2$ and $\Phi_I^\top r_{a,b} = y_1 + y_2$. Thus the margin in Eq. (3) becomes $\Delta_{a,b}^I(x) = \phi(x)^\top (G_1 + G_2)^{-1}(y_1 + y_2)$. The integrated model refits the readout rather than averaging source margins. This replacement of the source inverse-Gram operators by $(G_1 + G_2)^{-1}$ is inverse-Gram recomposition.

This change can reverse previously consistent preferences. A preference reversal occurs when, for the same action pair (a, b), models trained on individual sources prefer the same action while the integrated model prefers another action:

$$\Delta_{a,b}^1(x) > 0, \qquad \Delta_{a,b}^2(x) > 0, \qquad \Delta_{a,b}^I(x) < 0. \tag{4}$$

This condition means that both individual models prefer action a, while the integrated model prefers action b for the same decision problem. The reversal does not require conflicting information between sources. It can arise because information integration applies a different inverse-Gram operator to the combined source representations. Thus, information integration can change how a neural network interprets the same decision problem even when individual models initially agree.

Proposition 1 (Preference Reversal under Information Integration)**.** *There exist two finite sets of training cases with positive-definite Gram matrices for which* $\Delta_{a,b}^1(x) > 0$ *and* $\Delta_{a,b}^2(x) > 0$*, yet* $\Delta_{a,b}^I(x) < 0$ *after their cases are pooled and the readout is refitted.*

Proposition 1 establishes that source agreement alone cannot guarantee preservation. The explicit construction and its realization by case-representation matrices are given in Appendix B.1.

3.3 Preference Reversal and Decision Instability

Preference reversal describes a change in action preference, while decision instability occurs when such a change alters the final decision. For an input x, let $a_M(x)$ denote the action selected by a model trained on individual information sources, and let $a_I(x)$ denote the action selected by the integrated model after information integration. Since neural network decisions are obtained by selecting the highest-scoring action among candidates, a preference reversal affects the final decision only when it changes the selected action:

$$a_I(x) = \arg\max_a Q^I(x, a) \neq \arg\max_a Q^M(x, a) = a_M(x). \tag{5}$$

This condition defines decision instability and distinguishes it from preference reversal. Preference reversal describes a change in the relative preference between actions, whereas decision instability describes when this change modifies the final selected action. This distinction separates the mechanism causing instability from its final decision impact and explains how representation geometry influences decision stability.

4 Decision Stability Assurance before Information Integration

This section establishes how representation geometry can assure decision stability before information integration. We first derive a geometric condition that preserves decision preferences and then develop a measurable criterion for screening potential integration risks.

4.1 Geometric Alignment and Preference Preservation

This section derives the representation geometry condition that guarantees preference preservation during information integration. Let G_j be the Gram matrix of the cases from source j, evaluated under the shared representation map. Stable integration requires compatible source geometries, where aligned geometries share a common structure up to positive scaling:

$$G_j = \kappa_j G_0, \qquad \kappa_j > 0. \tag{6}$$

Here, G_0 denotes a reference geometry shared across sources, and κ_j captures source-specific scale differences. Under this condition, information integration preserves their common Gram direction. The superscript 12 is a label for sources 1 and 2 pooled together. The pooled preference margin is:

$$\Delta^{12}_{a,b}(x) = \frac{\kappa_1 \Delta^1_{a,b}(x) + \kappa_2 \Delta^2_{a,b}(x)}{\kappa_1 + \kappa_2}. \tag{7}$$

Proposition 2 (Preference preservation under aligned geometry)**.** *Under the shared fixed representation and full-rank OLS readout, suppose $G_j = \kappa_j G_0$ for $j \in \{1, 2\}$, with $G_0 \succ 0$ and $\kappa_j > 0$. The pooled margin is the positive weighted average in Eq. (7). If both source margins have the same nonzero sign, pooling preserves that pairwise preference.*

The derivation is given in Appendix B.2. Pairwise preference preservation also preserves a selected action when every comparison between that action and a competing action satisfies the condition.

4.2 Geometric Compatibility and Stability Conditions

This section develops a measurable geometry criterion for identifying whether candidate source models satisfy the stability condition before information integration. Different sources may induce different representation geometries. We quantify their geometric compatibility through normalized Gram mismatch:

$$\delta_{\text{norm}}(G_1, G_2) = \left\| \frac{G_1}{\|G_1\|_F} - \frac{G_2}{\|G_2\|_F} \right\|_F. \tag{8}$$

Here $\|\cdot\|_F$ is the Frobenius norm, which measures matrix magnitude. Normalizing by it removes positive scale differences. A smaller δ_{norm} means that the source Gram matrices have more similar directions.

Lemma 1 (Normalized Gram mismatch)**.** *For nonzero positive-semidefinite G_1, G_2, the mismatch is symmetric, invariant to independent positive rescaling, and lies in $[0, \sqrt{2}]$. It is zero exactly when G_1 and G_2 are positive scalar multiples.*

The proof is given in Appendix B.3.

To relate this criterion to decision stability, write each source Gram matrix as $G_j = \kappa_j G_0 + E_j$. Here E_j is its deviation from a shared reference matrix G_0, and κ_j is its positive scale. The pooled margin uses $(G_1 + G_2)^{-1}$, whereas the scale-weighted source margin uses $\kappa_j G_j^{-1}/(\kappa_1 + \kappa_2)$ for each source. We define ϵ_{geom} as an upper bound on the margin change caused by this difference:

$$\epsilon_{\text{geom}} = \|\phi(x)\|_2 \sum_{j=1}^{2} \left\| (G_1 + G_2)^{-1} - \frac{\kappa_j}{\kappa_1 + \kappa_2} G_j^{-1} \right\|_2 \|\Phi_j^\top d_j(a, b)\|_2. \tag{9}$$

Here $\|\cdot\|_2$ is the spectral norm for matrices and the Euclidean norm for vectors. The allowance is zero under scalar Gram alignment and increases with the difference between source and pooled inverse-Gram maps. It gives the following lower bound on the pooled margin:

$$\Delta^{12}_{a,b}(x) \geq \frac{\kappa_1 \Delta^1_{a,b}(x) + \kappa_2 \Delta^2_{a,b}(x)}{\kappa_1 + \kappa_2} - \epsilon_{\text{geom}}. \tag{10}$$

Theorem 1 (Preference preservation under approximate alignment)**.** *If the source Gram matrices are positive definite, the pooled margin obeys Eq. (10). When both sources favor the oriented action, a positive right-hand side certifies that their shared preference survives pooling.*

The proof and a perturbation bound on ϵ_{geom} are given in Appendix B.5. The conversion from normalized Gram mismatch to the deviation bound is derived in Appendix B.4. This links measurable mismatch to a sufficient margin condition for pre-integration screening, while retaining the source margins and query representation in the stability test.

5 Compositional Stability Enhancement during Information Integration

Representation geometry can be adjusted during representation learning to reduce incompatibility between information sources. We introduce geometry regularization to constrain learned representations and develop alignment objectives that improve compositional stability while maintaining task learning.

5.1 Representation Geometry Regularization

This section formulates representation geometry as a training objective for improving compositional stability during information integration. Let Φ_B denote the representation matrix of a training batch B, where each row corresponds to the learned representation of one sample. The induced Gram matrix G_B describes the geometry of the learned representations:

$$G_B = \Phi_B^\top \Phi_B. \tag{11}$$

Since representation scale does not determine geometric compatibility, we use the batch second-moment matrix and its normalized Gram direction:

$$\Psi_B = \frac{1}{|B|} G_B, \qquad \Gamma_B = \frac{\Psi_B}{\|\Psi_B\|_F} = \frac{G_B}{\|G_B\|_F}, \tag{12}$$

where Ψ_B is the batch second-moment matrix and Γ_B represents the scale-invariant geometry of the learned representations. Regulating these quantities during representation learning provides a direct way to reduce geometric differences between sources participating in information integration.

5.2 Geometry Alignment for Compositional Stability

This section introduces geometry alignment objectives that reduce representation incompatibility and improve compositional stability during information integration. We regulate representation geometry during representation learning together with the task objective. Let L_{task} denote the task loss and R_{geom} denote the geometry regularization term. The training objective is:

$$\mathcal{L} = L_{\text{task}} + \lambda_{\text{reg}} R_{\text{geom}}, \tag{13}$$

where λ_{reg} controls the contribution of geometry regularization. This objective balances task learning with the encouragement of representation geometries that are more compatible for information integration.

We first consider an isotropic reference geometry by reducing the deviation of the batch second moment from the identity:

$$R_{\text{white}} = \|\Psi_B - I_H\|_F^2, \tag{14}$$

where H is the representation width and I_H is its $H \times H$ identity matrix. This objective encourages different sources to approach a common representation geometry. We also consider direct alignment

between source representations. Let Γ_{ref} denote the normalized Gram geometry of the reference source. The alignment objective is:

$$R_{\mathrm{align}} = \|\Gamma_B - \Gamma_{\mathrm{ref}}\|_F^2, \tag{15}$$

This objective directly reduces the geometric difference between source representations. These two objectives reduce geometric incompatibility from complementary perspectives. By regulating representation geometry during representation learning, geometry alignment supports compositional stability when information from multiple sources is integrated.

6 Experiments

This section evaluates how representation geometry affects decision stability under information integration. We first examine whether geometric changes induce preference reversals, then analyze how these reversals affect decision outcomes and whether geometric compatibility enables assessment before information integration, and finally investigate whether geometry-oriented representation learning improves compositional stability.

We evaluate the theoretical findings in Sections 3–5 through experiments that examine preference reversal, geometric compatibility assessment, and compositional stability under information integration. In Section 6.1, we investigate whether representation geometry incompatibility can induce preference reversal using a controlled synthetic setting, and further examine whether this phenomenon occurs in an electricity-market decision task based on PJM Interconnection (PJM) historical load data (PJM Interconnection, 2017). In Section 6.2, we evaluate whether representation geometry compatibility can provide early compatibility assessment before information integration using the University of California, Irvine (UCI) Heart Disease and Default of Credit Card Clients datasets (Janosi et al., 1989; Yeh, 2009; Yeh & Lien, 2009). In Section 6.3, we study whether geometry regulation can improve compositional stability during information integration on the PJM load-based decision task. Detailed source construction, preprocessing, task-defined returns, training settings, and statistical units are provided in Appendix C.

6.1 Analyzing Representation Geometry and Preference Reversals

We investigate whether representation geometry changes can disrupt previously established preferences after information integration. We first examine this mechanism under a controlled setting and then evaluate whether the same phenomenon appears in a PJM load-based decision task.

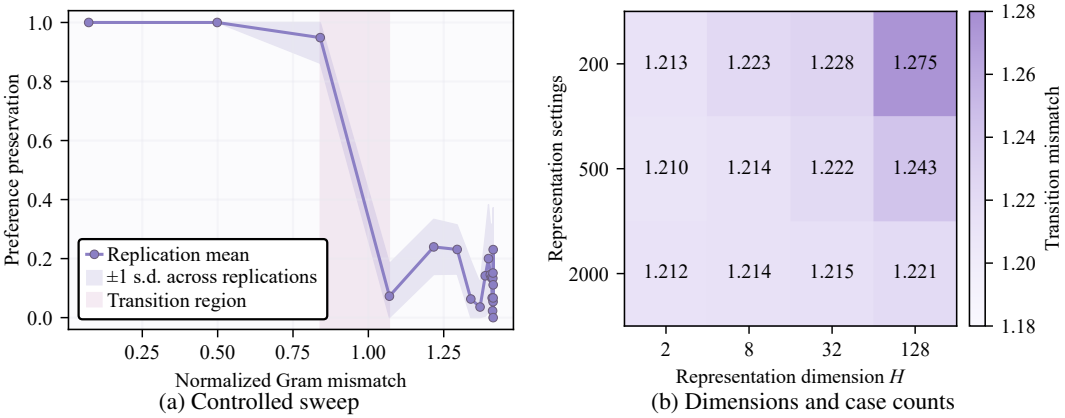


Figure 2: Controlled study of how representation geometry mismatch affects preference preservation during information integration. (a) Preference preservation remains high under small normalized Gram mismatch but decreases rapidly after entering the transition region where preference reversals emerge. (b) The transition region appears within a similar mismatch range across different representation settings, indicating that normalized Gram mismatch identifies the boundary between stable and unstable integration.

Table 1: Preference reversals and their downstream decision outcomes on the PJM load-based decision task. Among eligible query–action pairs, 1.80% exhibit preference reversal after integration. Among source-preference-agreement queries, 1.12% lead to changed decisions, and 53.64% of changed decisions result in utility decreases. These outcomes are measured using task-defined returns.

Evaluation stage	Evaluation population	Result
Preference reversal	Eligible query–action pairs	**626 / 34,693 (1.80%)**
Decision change	Source-consensus queries	**110 / 9,837 (1.12%)**
Utility decrease	Changed decisions	**59 / 110 (53.64%)**

The controlled results reveal a clear separation between geometrically compatible and incompatible integrations. As shown in Figure 2(a), when source representations remain close in geometry, the integrated model preserves the preference ordering supported by individual models. However, after geometric mismatch enters the transition region, the integrated model begins to produce different preference rankings. Since the source models maintain consistent preferences before integration, this result rules out conflicting source decisions as an explanation and indicates that the change emerges during representation recomposition. As shown in Figure 2(b), the transition region appears within a similar mismatch range across different representation settings, indicating that the observed reversal behavior is not specific to a particular configuration. These results show that representation geometry determines whether individual model preferences can be preserved after information integration, supporting the theoretical mechanism that geometric incompatibility can induce preference reversal. Additional threshold sensitivity and a representative reversal record are provided in Appendix D.

The PJM load-based decision task further shows that preference reversal is not limited to controlled perturbations but also appears in the load-based decision task. As shown in Table 1, preference reversal occurs at the preference level, while decision changes are observed under source preference agreement and may further lead to utility decreases. This separation indicates that preference instability and decision instability correspond to different stages of integration behavior. Moreover, the fact that only some changed decisions reduce utility shows that preference reversal reflects a loss of preference preservation rather than a deterministic degradation of task utility. Together with the controlled experiment, these results support the theoretical mechanism that representation geometry changes during information integration can alter previously established preferences.

6.2 Assessing Geometric Compatibility before Information Integration

We evaluate whether representation geometry compatibility can provide an early compatibility assessment before information integration.

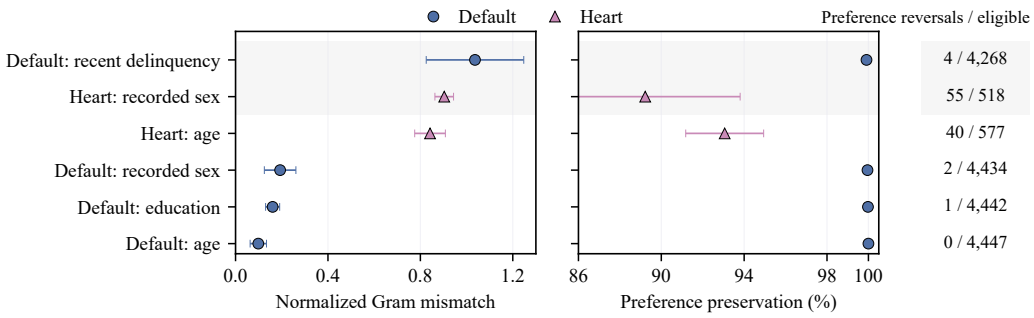


Figure 3: Representation geometry compatibility and preference preservation across datasets. Source representation combinations with different geometric compatibility exhibit different preference preservation outcomes across the Heart Disease and Default of Credit Card Clients datasets. Larger normalized Gram mismatch is associated with more observed preference reversals, while lower mismatch corresponds to more stable preference preservation.

Table 2: Comparison of geometry regulation strategies for improving compositional stability under information integration. Gram whitening reduces normalized Gram mismatch by 10.9% and improves preference preservation over unregularized training, while maintaining a comparable utility-decrease rate to alternative regulation strategies. Here, h_{all} denotes the overall utility-decrease rate.

Method	Accuracy ↑	Preservation ↑	δ_{norm} ↓	$h_{\text{all}}(\%)$ ↓
No regularization	50.05±10.13	98.16±1.10	1.198±0.056	0.35±0.12
Gram whitening, 0.01	**52.98±3.73**	**98.83±0.25**	**1.068±0.051**	0.40±0.23
Direct alignment, 0.1	49.84±10.04	98.20±1.05	1.196±0.056	**0.31±0.05**
Weight decay, 0.01	50.98±5.47	97.64±1.19	1.168±0.074	0.81±0.90

Experiments on the Heart Disease and Default of Credit Card Clients datasets show that representation geometry mismatch is associated with differences in preference preservation after information integration (Figure 3). Across different source representation combinations, larger geometric incompatibility corresponds to less stable preference preservation. This consistent pattern across datasets suggests that representation geometry captures a general factor affecting integration stability rather than a dataset-specific property. Therefore, representation geometry provides an informative criterion for assessing potentially unstable information integration before information integration, supporting the theoretical analysis in Section 4. Complete source-partition counts and preservation results are reported in Appendix F.

6.3 Evaluating Geometry Regulation for Compositional Stability

We investigate whether regulating representation geometry during representation learning can improve compositional stability under information integration.

Geometry regulation improves the compositional stability of integrated models by controlling the compatibility between source representations. As shown in Table 2, Gram whitening reduces normalized Gram mismatch and improves preference preservation compared with unregularized training, indicating that explicitly regulating representation geometry helps preserve source preferences during information integration. Although direct alignment achieves a slightly lower utility-decrease rate, Gram whitening provides stronger improvements in representation compatibility and preference preservation, highlighting the role of geometry regulation in maintaining stable information integration. These results support the theoretical analysis in Section 5 that representation geometry can be optimized during learning to improve the stability of integrated decisions. The complete configuration sweep and paired uncertainty estimates are provided in Appendix E.

7 Conclusion

Information integration can alter previously established decision preferences because refitting on pooled information changes the representation geometry used by the integrated model. This work identifies representation geometry as a key factor determining whether source preferences are preserved during information integration. Through theoretical analysis, we show that geometric incompatibility can induce preference reversal by changing the transformation applied during integration, and derive conditions under which source preferences can be preserved. The experiments further demonstrate that preference reversal can emerge in both controlled settings and experiments using real-world datasets, that geometric compatibility provides an effective criterion for assessing potentially unstable information integration before information integration, and that geometry regulation during representation learning can improve compositional stability. Together, these results establish a unified framework for understanding, assessing, and improving compositional stability under information integration in multi-source decision systems.

This work focuses on fixed representations with OLS readouts, and extending the analysis to jointly optimized representations and nonlinear decision heads remains an important direction. Future research can further investigate how geometry regulation can be integrated into broader representation learning frameworks to improve compositional stability in more complex information integration settings.

AI USE STATEMENT

AI tools assisted with polishing manuscript language, language translation, searching for related literature, and document checks. The authors manually verified the literature identified through AI-assisted search. All other research content and activities, including the research ideas, method development, theoretical derivations, experimental design and execution, and analysis of the results, were completed by the authors. The authors take responsibility for the content of this paper.

ETHICS STATEMENT

This study uses public PJM load data, Heart Disease, and Default of Credit Card Clients for computational evaluation. It involves no new collection of participant data or animal experiments. Demographic and financial attributes are used to construct source partitions for studying information integration. The action returns are constructed for methodological evaluation, and preference stability does not by itself establish clinical utility, lending suitability, or fairness. The analysis uses the released records without attempting to identify individuals.

REPRODUCIBILITY STATEMENT

Appendix B states the assumptions and provides the proofs of the theoretical results. Appendix C describes the datasets, source partitions, preprocessing, task-defined returns, network configurations, training objectives, and statistical units. Appendices D–F provide additional results and uncertainty estimates. The experiment files, fixed splits, recorded outputs, and deterministic reproduction script are described in Appendix C.6. The public data sources are cited in Section 6.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. doi: 10.1007/s10994-009-5152-4.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://jmlr.org/papers/v17/15-239.html.

Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html.

Itzhak Gilboa and David Schmeidler. Case-based decision theory. *The Quarterly Journal of Economics*, 110(3):605–639, 1995. doi: 10.2307/2946694.

Itzhak Gilboa, David Schmeidler, and Peter P. Wakker. Utility in case-based decision theory. *Journal of Economic Theory*, 105(2):483–502, 2002. doi: 10.1006/jeth.2001.2858. URL https://www.sciencedirect.com/science/article/abs/pii/S0022053101928583.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Huan He, Owen Queen, Teddy Koker, Consuelo Cuevas, Theodoros Tsiligkaridis, and Marinka Zitnik. Domain adaptation for time series under feature and label shifts. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12746–12774, 2023.

Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 791–800, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Huang_Decorrelated_Batch_Normalization_CVPR_2018_paper.html.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1989. URL https://archive.ics.uci.edu/dataset/45/heart+disease. Dataset. Accessed March 22, 2026.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Xinyao Li, Jingjing Li, Fengling Li, Lei Zhu, and Ke Lu. Agile multi-source-free domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13673–13681, 2024. doi: 10.1609/aaai.v38i12.29272.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/ab88b15733f543179858600245108dd8-Abstract.html.

Na Luo, Weiyang Shi, Zhengyi Yang, Ming Song, and Tianzi Jiang. Multimodal fusion of brain imaging data: Methods and applications. *Machine Intelligence Research*, 21:136–152, 2024. doi: 10.1007/s11633-023-1442-8.

Michael S. Matena and Colin A. Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/70c26937fbf3d4600b69a129031b66ec-Abstract-Conference.html.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5727–5736, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a7a3d70c6d17a73140918996d03c014f-Abstract.html.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=KJNcAkY8tY4.

PJM Interconnection. Historical Metered Load Data. PJM Data Directory, 2017. URL https://www.pjm.com/markets-and-operations/data-dictionary.aspx?p=1. 2017 load data. Accessed July 12, 2023.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6076–6085, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 2016. doi: 10.1609/aaai.v30i1.10306. URL https://ojs.aaai.org/index.php/AAAI/article/view/10306.

Lisheng Wen, Sentao Chen, Mengying Xie, Cheng Liu, and Lin Zheng. Training multi-source domain adaptation network by mutual information estimation and minimization. *Neural Networks*, 171: 353–361, 2024. doi: 10.1016/j.neunet.2023.12.022.

Tom N. Wolf, Fabian Bongratz, Anne-Marie Rickmann, Sebastian Pölsterl, and Christian Wachinger. Keep the faith: Faithful explanations in convolutional neural networks for case-based reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5921–5929, 2024. doi: 10.1609/aaai.v38i6.28406.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Manli Yan, Yuebin Lin, Yaowen Yu, and Yong Zhao. From neural network decisions to training cases: An exact account via case-based decision theory, 2026. URL https://arxiv.org/abs/2607.11347.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9291–9301, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/8a7129b8f3edd95b7d969dfc2c8e9d9d-Abstract.html.

I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. URL https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients. Dataset. Accessed March 22, 2026.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2): 2473–2480, 2009. doi: 10.1016/j.eswa.2007.12.020.

Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9):1–36, 2024. doi: 10.1145/3649447. Article 216.

Han Zhao, Shanghang Zhang, Guanhang Wu, Jose M. F. Moura, Joao P. Costeira, and Geoffrey J. Gordon. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/717d8b3d60d9eea997b35b02b6a4e867-Abstract.html.

A NOTATION

Table 3: Notation for preference preservation, Gram stability, and the experiments.

Symbol	Meaning
x	Test query.
$\mathcal{A}, K$	Finite action set and number of actions.
a, b	Candidate actions in an audited pair.
M_j, M	Source model j and generic fitted model.
$\mathcal{X}_{\text{test}}$	Test-query set.
$\phi(x), \Phi, \Phi_j, \Phi_I$	Query representation and generic, source, and integrated representation matrices.
$Q^M(x,a), \widehat{v}_a^M, v^M$	OLS score, fitted action coefficient, and coefficient difference.
$r_{a,b}, d_j(a,b)$	Generic and source action-return differences.
y_j	Source evidence vector $\Phi_j^\top d_j(a,b)$.
$G, G_j, G_I, \widetilde{G}_j, G_0$	Generic, source, integrated, normalized, and reference Gram matrices.
δ_{norm}	Scale-invariant normalized Gram mismatch.
y	Pair-evidence vector in the ridge perturbation bound.
$\Delta_{a,b}^M, \Delta_{a,b}^I, \Delta_{a,b}^j, \Delta_{a,b}^{12}$	Generic, integrated, source, and pooled margins for action pair (a,b).
$m_0, m_{\lambda_{\text{ridge}}}$	OLS and ridge-stabilized margins.
$\lambda_{\text{ridge}}, \mu$	Ridge stabilization and $\lambda_{\min}(G)$.
τ	Minimum shared source-margin threshold in the audit.
$\epsilon_{\text{pair}}, \epsilon_{\text{tie}}$	Pairwise and decision-tie tolerances.
$a_M, a_I, a_j, a_{12}, a_s$	Generic-model, integrated, source, pooled, and source-consensus actions.
C, D	Source-consensus and decision-change sets.
$u(x,a), \ell_u(x)$	Task utility and signed utility loss.
h_D, h_{all}, h_C	Changed-decision, overall, and consensus utility-loss rates.
κ_j, κ_Σ	Source Gram scale and $\kappa_1 + \kappa_2$.
$\kappa_{\min}, \kappa_{\max}$	Minimum and maximum source Gram scales.
$E_j, \epsilon, \epsilon_\delta, \epsilon_{\text{geom}}$	Gram perturbation and perturbation levels.
λ_0	Minimum eigenvalue of the reference Gram matrix G_0.
G_Σ, H_j, u_j	Pooled proof matrix $G_1 + G_2$, rescaled source matrix $(\kappa_\Sigma/\kappa_j)G_j$, and reference-geometry evidence projection.
$\|\cdot\|_F, \|\cdot\|_2$	Frobenius and Euclidean/spectral norms.
B, Φ_B, G_B, I_H	Mini-batch, batch representation matrix, batch Gram matrix, and identity.
Ψ_B, Γ_B	Batch second moment and its normalized Gram direction.
Γ_{ref}	Reference normalized Gram direction from the lower-demand stage.
$\mathcal{L}, L_{\text{task}}, R_{\text{geom}}, \lambda_{\text{reg}}$	Total objective, task loss, geometry regularizer, and strength.
$R_{\text{white}}, R_{\text{align}}$	Whitening and direct-alignment penalties.
N, α, H	Controlled source case count, anisotropy strength, and dimension.
c	Dataset class or target label for task-defined returns.
i, h	Ordered load row/day index and half-hour slot.
$L_{i,h}, \bar{L}_{i,h}$	Observed and clipped normalized load.
$\rho_{i,h}, m_{i,h}$	Load ramp and market price.
$z_{i,h}^{(0)}, \tilde{z}_{i,h}$	Initial and adjusted deterministic price components.
$\xi_{i,h}, \eta_{i,h}$	Gaussian price noise and rare spike component.
$b_a, r(a,m)$	Bid level and task-defined bidding return.
Acc.	Fraction of test queries for which the pooled action equals the task-defined oracle action.
$\|D\|/\|\mathcal{X}_{\text{test}}\|$	Overall decision-change rate.
$\mathbf{1}\{\cdot\}, q$	Indicator function and generic action-score vector.
$L, f_\theta, r(x)$	Utility-loss set, temporary action-value head, and target return vector.
θ, β_1, β_2	Representation parameters and Adam coefficients.
s, p	Source-consensus and pooled actions in the representative reversed pair.

B PROOFS AND DERIVATIONS

B.1 REVERSAL CONSTRUCTION

Let

$$G_1 = \begin{bmatrix} 4.022 & 1.044 \\ 1.044 & 0.631 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 0.644 & -0.578 \\ -0.578 & 2.451 \end{bmatrix}, \\ \phi(x) = \begin{bmatrix} 1.442 \\ 1.021 \end{bmatrix}, \quad y_1 = \begin{bmatrix} -0.750 \\ 0.214 \end{bmatrix}, \quad y_2 = \begin{bmatrix} 0.267 \\ -0.209 \end{bmatrix}. \tag{B1}$$

The eigenvalues of G_1 are 0.335 and 4.318, and those of G_2 are 0.475 and 2.620, so both matrices are positive definite. Direct substitution gives $\phi(x)^\top G_1^{-1} y_1 = 0.465$, $\phi(x)^\top G_2^{-1} y_2 = 0.635$, and $\phi(x)^\top (G_1 + G_2)^{-1}(y_1 + y_2) = -0.134$. Thus both sources favor the oriented action pair, whereas the pooled readout reverses the sign. These matrices can be realized by finite sets of cases. For each $G_j \succ 0$, choose a full-column-rank matrix Φ_j satisfying $\Phi_j^\top \Phi_j = G_j$, and set $d_j = \Phi_j G_j^{-1} y_j$. Then $\Phi_j^\top d_j = y_j$. Thus the source margins arise from valid representation and return-difference matrices.

B.2 GUARANTEED PRESERVATION UNDER SCALAR GRAM ALIGNMENT

Suppose $G_j = \kappa_j G_0$ with $G_0 \succ 0$ and $\kappa_j > 0$. Then $(G_1 + G_2)^{-1} = (\kappa_1 + \kappa_2)^{-1} G_0^{-1}$ and $G_j^{-1} = \kappa_j^{-1} G_0^{-1}$. With $u_j = \phi(x)^\top G_0^{-1} y_j$, we have $u_j = \kappa_j \Delta_{a,b}^j(x)$ and therefore

$$\Delta_{a,b}^{12}(x) = \frac{u_1 + u_2}{\kappa_1 + \kappa_2} = \frac{\kappa_1 \Delta_{a,b}^1(x) + \kappa_2 \Delta_{a,b}^2(x)}{\kappa_1 + \kappa_2}. \tag{B2}$$

Both weights are positive and sum to one. A pairwise preference held weakly by both sources therefore survives pooling. If either source prefers the oriented action strictly, the pooled margin is strictly positive.

B.3 PROPERTIES OF NORMALIZED GRAM MISMATCH

For nonzero positive-semidefinite Gram matrices, let $\widetilde{G}_j = G_j / \|G_j\|_F$. Then $\|\widetilde{G}_j\|_F = 1$ and $\langle \widetilde{G}_1, \widetilde{G}_2 \rangle_F = \mathrm{tr}(\widetilde{G}_1 \widetilde{G}_2) \geq 0$. Consequently,

$$\delta_{\mathrm{norm}}^2(G_1, G_2) = 2 - 2\langle \widetilde{G}_1, \widetilde{G}_2 \rangle_F \leq 2. \tag{B3}$$

This distance is symmetric and unchanged by independent positive rescaling. It is zero exactly when the two raw matrices are positive scalar multiples. If both matrices are positive definite, then $\mathrm{tr}(\widetilde{G}_1 \widetilde{G}_2) > 0$ and the upper bound is strict: $\delta_{\mathrm{norm}}(G_1, G_2) < \sqrt{2}$.

B.4 FROM OBSERVED MISMATCH TO A COMMON PERTURBATION

Let $\kappa_j = \|G_j\|_F$ and choose $G_0 = (\widetilde{G}_1 + \widetilde{G}_2)/2$ and $E_j = G_j - \kappa_j G_0$. Because $G_1, G_2 \succ 0$, their normalized average satisfies $G_0 \succ 0$. Moreover,

$$E_1 = \frac{\kappa_1}{2}(\widetilde{G}_1 - \widetilde{G}_2), \qquad E_2 = \frac{\kappa_2}{2}(\widetilde{G}_2 - \widetilde{G}_1), \tag{B4}$$

which gives

$$\|E_j\|_2 \leq \|E_j\|_F = \frac{\kappa_j}{2} \delta_{\mathrm{norm}}(G_1, G_2). \tag{B5}$$

This decomposition makes the perturbation size observable from the source Gram matrices. Define $\epsilon_\delta = \delta_{\mathrm{norm}}(G_1, G_2)\kappa_{\max}/(2\|G_0\|_2)$, where $\kappa_{\max} = \max_j \kappa_j$. Then $\|E_j\|_2 \leq \epsilon_\delta \|G_0\|_2$ for both sources. Since $\kappa_\Sigma \geq 2\kappa_{\min}$, the smaller term in Eq. (B6) is $\kappa_{\min}/2$. Therefore the directly checkable condition $\delta_{\mathrm{norm}}(G_1, G_2) < \lambda_0 \kappa_{\min}/\kappa_{\max}$ implies Eq. (B6) after substituting ϵ_δ.

B.5 Approximate-alignment bound

Let $G_j = \kappa_j G_0 + E_j$ for $j \in \{1,2\}$, where $G_0 \succ 0$, $\kappa_j > 0$, and $\|E_j\|_2 \le \epsilon\|G_0\|_2$. Define $\kappa_{\min} = \min\{\kappa_1, \kappa_2\}$, $\kappa_\Sigma = \kappa_1 + \kappa_2$, $\lambda_0 = \lambda_{\min}(G_0)$, and $y_j = \Phi_j^\top d_j(a,b)$. Assume

$$\epsilon\|G_0\|_2 < \lambda_0 \min\left\{\frac{\kappa_{\min}}{2}, \frac{\kappa_\Sigma}{4}\right\}. \tag{B6}$$

We show that

$$\begin{aligned}\Delta^{12}_{a,b}(x) \ge{}& \frac{\kappa_1\Delta^1_{a,b}(x) + \kappa_2\Delta^2_{a,b}(x)}{\kappa_\Sigma} \\ &- \frac{4\epsilon\|\phi(x)\|_2\|G_0\|_2}{\kappa_\Sigma\lambda_0^2}\left(\frac{\|y_1\|_2}{\kappa_1} + \frac{\|y_2\|_2}{\kappa_2}\right).\end{aligned} \tag{B7}$$

The second term in Eq. (B7) is an explicit upper bound on the geometry allowance in Eq. (10): $\epsilon_{\text{geom}} \le 4\epsilon\|\phi(x)\|_2\|G_0\|_2(\|y_1\|_2/\kappa_1 + \|y_2\|_2/\kappa_2)/(\kappa_\Sigma\lambda_0^2)$. The first term is the pooled margin under Gram alignment. The second bounds how far the actual margin can fall below it. If their difference is positive, the oriented source preference survives pooling.

To prove Eq. (B7), set $G_\Sigma = G_1 + G_2$ and $H_j = (\kappa_\Sigma/\kappa_j)G_j$. Then $H_j^{-1} = \kappa_j G_j^{-1}/\kappa_\Sigma$ and $\phi(x)^\top H_j^{-1} y_j = \kappa_j\Delta^j_{a,b}(x)/\kappa_\Sigma$. The condition in Eq. (B6) gives

$$\lambda_{\min}(G_\Sigma) \ge \kappa_\Sigma\lambda_0 - 2\epsilon\|G_0\|_2 > \frac{\kappa_\Sigma\lambda_0}{2}, \tag{B8}$$

and, because $\kappa_j \ge \kappa_{\min}$,

$$\lambda_{\min}(H_j) \ge \kappa_\Sigma\lambda_0 - \frac{\kappa_\Sigma}{\kappa_j}\epsilon\|G_0\|_2 > \frac{\kappa_\Sigma\lambda_0}{2}. \tag{B9}$$

Thus $\|G_\Sigma^{-1}\|_2$ and $\|H_j^{-1}\|_2$ are both at most $2/(\kappa_\Sigma\lambda_0)$. In addition,

$$H_j - G_\Sigma = \frac{\kappa_{3-j}}{\kappa_j}E_j - E_{3-j}, \qquad \|H_j - G_\Sigma\|_2 \le \frac{\kappa_\Sigma}{\kappa_j}\epsilon\|G_0\|_2. \tag{B10}$$

Using the resolvent identity $G_\Sigma^{-1} - H_j^{-1} = G_\Sigma^{-1}(H_j - G_\Sigma)H_j^{-1}$ yields

$$\|G_\Sigma^{-1} - H_j^{-1}\|_2 \le \frac{4\epsilon\|G_0\|_2}{\kappa_\Sigma\kappa_j\lambda_0^2}. \tag{B11}$$

Since $H_j^{-1} = \kappa_j G_j^{-1}/\kappa_\Sigma$, applying Eq. (B11) to the definition of ϵ_{geom} yields the stated perturbation bound. Cauchy–Schwarz also gives

$$\phi(x)^\top G_\Sigma^{-1} y_j \ge \frac{\kappa_j\Delta^j_{a,b}(x)}{\kappa_\Sigma} - \frac{4\epsilon\|\phi(x)\|_2\|G_0\|_2\|y_j\|_2}{\kappa_\Sigma\kappa_j\lambda_0^2}. \tag{B12}$$

Summing Eq. (B12) over $j = 1, 2$ and using $\Delta^{12}_{a,b}(x) = \phi(x)^\top G_\Sigma^{-1}(y_1 + y_2)$ proves Eq. (B7).

B.6 Ridge-stabilized margin perturbation

For $G \succ 0$ and $\lambda_{\text{ridge}} \ge 0$, let $m_0 = \phi(x)^\top G^{-1} y$, $m_{\lambda_{\text{ridge}}} = \phi(x)^\top (G + \lambda_{\text{ridge}} I)^{-1} y$, and $\mu = \lambda_{\min}(G) > 0$. The resolvent identity gives

$$(G + \lambda_{\text{ridge}} I)^{-1} - G^{-1} = -(G + \lambda_{\text{ridge}} I)^{-1}\lambda_{\text{ridge}} G^{-1}. \tag{B13}$$

Since $\|(G + \lambda_{\text{ridge}} I)^{-1}\|_2 \le (\mu + \lambda_{\text{ridge}})^{-1}$ and $\|G^{-1}\|_2 = \mu^{-1}$, Cauchy–Schwarz implies

$$|m_{\lambda_{\text{ridge}}} - m_0| \le \frac{\lambda_{\text{ridge}}\|\phi(x)\|_2\|y\|_2}{\mu(\mu + \lambda_{\text{ridge}})}. \tag{B14}$$

Whenever $|m_0|$ exceeds this bound, the ridge margin has the same sign as the OLS margin. The condition concerns a fixed Gram matrix and evidence vector. It does not remove the need to check compatibility between sources.

C EXPERIMENTAL PROTOCOL

This section specifies the observations, source partitions, task-defined returns, and replication units underlying the main results. Representation widths below refer to rectified linear unit (ReLU) layers.

Table 4: Experimental roles. Each row identifies the source construction, action or task-defined return, statistical unit, and scientific question assigned to that study.

Experiment	Source construction	Actions/returns	Primary unit	Evidence role
Controlled	Synthetic Gram-direction sweep	Three-action fixed protocol	Eligible pair	Reversal mechanism and transition location
Electricity market bidding task	Training-defined low/high load case sets	Three bidding-task returns	Pair, consensus test query, outcome	Full consequence chain
Heart Disease	Age or recorded-sex partitions	Triage-intensity task	Eligible pair	Cross-domain scope
Default Credit	Age, education, recorded sex, recent delinquency	Exposure task	Eligible pair	Cross-domain screening scope
Operating points	Matched load runs across objectives	Same per-run returns	Matched run with pair/test-query summaries	Geometry, preservation, and utility

Table 5: Sampling and model settings for each domain. Table 4 gives the corresponding statistical units and experimental roles.

Domain	Replications and split	Repre-sentation	Source rule and size	Task returns and ridge
Controlled	Five replications with an accepted-query test stream	H varies	Two source sets of 600 cases each. Accepted queries require both source directions to support the evaluated action contrast	Three-action fixed return protocol with ridge 10^{-6}
Load	Five matched runs with a chronological 13,432/3,358 split	32-ReLU	Training-median current load. Lower/higher source sizes 6,717/6,715 in every run	Bids 25/45/70 USD/MWh with price-dependent returns with ridge 10^{-5}
Heart Disease	Five random 80/20 splits with 237/60 per split	64-ReLU	Training-median age or recorded sex. Age sources 112–116/121–125, sex sources 156–162/75–81	Three triage-intensity task-defined returns with ridge 10^{-5}
Default Credit	Five random 80/20 splits with 24,000/6,000 per split	64–32–16	Median age, education, recorded sex, or recent delinquency. Source-size ranges are 5,376–19,748 and 4,252–18,624 across partitions	Three exposure task-defined returns with ridge 10^{-5}

C.1 PREFERENCE PRESERVATION AND DECISION AUDIT

The canonical audit sets $\tau = 0$ and $\epsilon_{\text{pair}} = 10^{-6}$. For each test query it enumerates all $\binom{K}{2}$ unordered action pairs and retains the pairs whose two source margins share the same nonzero sign. The audit counts strict reversals and pooled ties separately. For each replication, the strict preference-preservation rate is the fraction of eligible pairs for which the pooled margin retains the common source sign. Pooled ties count as failures. These rates concern pairs with strict agreement between source preferences. Appendix D.2 increases τ to measure how preservation changes as the shared source margin becomes stronger. Weak–strict cases form a separate extension.

For completeness, the deterministic decision rule is

$$\operatorname*{argmax}_{\epsilon_{\text{tie}}} q = \min\{a : q_a \geq \max_b q_b - \epsilon_{\text{tie}}\}, \qquad \epsilon_{\text{tie}} = 10^{-6}.$$

On $C = \{x : a_1(x) = a_2(x)\}$, let $a_s(x) = a_1(x)$ and $D = \{x \in C : a_{12}(x) \neq a_s(x)\}$. Define

$$\ell_u(x) = u(x, a_s(x)) - u(x, a_{12}(x)), \qquad L = \{x \in D : \ell_u(x) > 0\}.$$

The utility-loss fractions are $h_D = |L|/|D|$, $h_C = |L|/|C|$, and $h_{\text{all}} = |L|/|\mathcal{X}_{\text{test}}|$, whenever the corresponding denominator is nonzero. We count $\ell_u(x) < 0$ as a gain and $\ell_u(x) = 0$ as neutral. To associate a changed decision with a strict reversal, we orient its action pair as $(a_s(x), a_{12}(x))$ and require both source margins to exceed τ and the pooled margin to be below $-\epsilon_{\text{pair}}$.

C.2 Controlled construction and dimension/memory-size grid

The main controlled sweep varies $\alpha \in \{0, 0.05, 0.10, \ldots, 0.95, 1.00\}$, which changes the two source feature scales to $(e^{-2.5\alpha}, e^{2.5\alpha})$ and $(e^{2.5\alpha}, e^{-2.5\alpha})$. The return-difference directions remain fixed at $d_1 = (3, -1)$ and $d_2 = (-1, 3)$, with $\phi(x) = x$. A candidate query enters the test set when both source directions give nonnegative support to the evaluated contrast. In source j, action 0 has base return $\phi(x)^\top d_j$, while actions 1 and 2 start at zero before independent Gaussian noise is added. Memory generation and accepted-query generation use separate pseudorandom streams within each replication.

The main sweep adds Gaussian return noise with standard deviation 0.05 and uses two 600-case source sets, three actions, 800 accepted test queries per replication, five replications per setting, and predictor ridge 10^{-6}. The robustness grid varies $H \in \{2, 8, 32, 128\}$ and $N \in \{200, 500, 2000\}$, with 300 accepted test queries, $\alpha \in \{0, 0.1, \ldots, 1.0\}$, return-noise standard deviation 0.3, and five replications per cell. Its transition statistic is the first sampled Gram mismatch at which replication-mean strict preference-preservation rate falls below 0.5. The reported value is one of the evaluated grid points.

C.3 Task-defined returns and source partitions

For Heart Disease, we use the 297 complete Cleveland records and all 13 published attributes. Five independent unstratified 80/20 splits give 237 training and 60 test records per run. Continuous variables are standardized with training statistics, while categorical variables keep their published numeric codes. Each training split is partitioned by median age or recorded sex. Diagnosis class $c \in \{0, \ldots, 4\}$ maps to the three-action task-defined return vectors $(1.0, 0.3, -0.2)$, $(0.2, 0.9, 0.5)$, $(-0.1, 1.0, 0.7)$, $(-0.3, 0.6, 1.0)$, and $(-0.5, 0.4, 1.1)$.

Default Credit also uses five independent unstratified 80/20 splits. The identifier and target are removed from the inputs, published numeric category codes are retained, and all standardization constants and partition thresholds are fitted on the training split. Source partitions use training-median age, education (EDUCATION ≤ 2 versus > 2), recorded sex, or recent delinquency (PAY_0 > 0 versus ≤ 0). For default indicator $c \in \{0, 1\}$, the three exposure-task-defined returns are $(0.5(1-c), 1-2c, 2-5c)$. These returns define the decision task for the composition audit, and recorded sex serves as one source partition in the methodological stress test.

C.4 Matched load returns and utility

Each matched load run contains 16,790 ordered row–slot queries, with 13,432 chronological training queries and 3,358 test queries. Across five runs, the staged audit therefore contains 16,790 test-query evaluations. Observed load supplies the covariates. Synthetic prices, bids, and action returns provide a complete utility table so that the consequence of every selected action can be evaluated.

The archived PJM array has 365 ordered rows and 48 half-hour slots per row. The artifact contains the ordered array used in the experiments, preprocessing code, run-specific synthetic-return generators, split indices, and result manifests. The cited PJM Data Directory documents the public load-data family. Because the feature vector uses a two-slot lag, the analysis retains slots $h = 2, \ldots, 47$. The experiments use the archived array in its stored order, without interpolation, duplication, or resampling. The training-period median current load defines the two source case sets, giving 6,717 lower-demand and 6,715 higher-demand training queries in every matched run.

For row–slot index (i, h), the four-dimensional input is

$$x_{i,h} = (L_{i,h}, L_{i,h-2}, \sin(2\pi h/48), \cos(2\pi h/48)) .$$

Continuous load features are standardized with statistics from the chronological training period before representation training. The fitted transformation is then applied unchanged to both source case sets and the held-out test queries.

The clipped normalized load is

$$\bar{L}_{i,h} = \mathrm{clip}\left(\frac{L_{i,h} - 700}{600}, -0.2, 1.5\right),$$

and the deterministic price component before time and ramp adjustments is

$$z^{(0)}_{i,h} = 28 + 18\bar{L}_{i,h} + 22\bar{L}^2_{i,h}.$$

Let $\tilde{z}_{i,h}$ be the deterministic price after peak and ramp adjustments. Starting from $z^{(0)}_{i,h}$, we multiply by 1.12 for $28 \le h \le 38$. With ramp $\rho_{i,h} = (L_{i,h} - L_{i,h-2})/L_{i,h-2}$, values above 0.04 receive the additional factor $1 + 1.8(\rho_{i,h} - 0.04)$. A matched run then adds Gaussian noise $\xi_{i,h} \sim \mathcal{N}(0, (0.18\tilde{z}_{i,h})^2)$ and, with probability 1% when load exceeds 1100, a scarcity spike $\eta_{i,h}$ drawn uniformly from 50 to 150. The persisted constructed market-clearing price is $m_{i,h} = \max\{15, \tilde{z}_{i,h} + \xi_{i,h} + \eta_{i,h}\}$.

The bid levels are $b = (25, 45, 70)$ USD/MWh. For market price m, the bidding-task return is

$$r(a, m) = \begin{cases} (m - b_a)(1 + 0.6a) - 18, & m \ge b_a, \\ -3(a + 1), & m < b_a. \end{cases}$$

The persisted return table defines $u(x, a)$ for every candidate action at every training and test query. Utility loss or gain is computed directly from the difference between the pooled action's return and the source-consensus action's return. Exact equality is neutral.

Each run uses a separate pseudorandom stream. The price and action-return arrays are saved before the methods are compared, so all configurations within a run use identical utilities.

C.5 Training objectives and statistical units

Each method/run combination produces one representation map ϕ, shared by every source and pooled audit readout in that run. The same map determines the source and pooled Gram matrices. During representation training, a temporary action-value head minimizes $L_{\mathrm{task}} = |B|^{-1} \sum_{(x,r)\in B} \|f_\theta(x) - r(x)\|_2^2$. This head is discarded before the audited OLS/ridge readouts are fitted.

The load representation has one 32-unit ReLU hidden layer. Heart Disease uses one 64-unit ReLU layer, and Default Credit uses hidden widths 64–32–16. All representations use Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, learning rate 10^{-3}, batch size 128, and 80 epochs per stage. Geometry penalties enter as $L_{\mathrm{task}} + \lambda_{\mathrm{reg}} R$, with $R_{\mathrm{white}}(B) = \|\Psi_B - I_H\|_F^2$ or $R_{\mathrm{align}}(B) = \|\Gamma_B - \Gamma_{\mathrm{ref}}\|_F^2$. Direct alignment is applied in the second, higher-demand stage using the terminal first-stage Gram direction as its reference, so the procedure depends on source order.

For the electricity-market bidding task, decision accuracy is $\mathrm{Acc} = |\mathcal{X}_{\mathrm{test}}|^{-1} \sum_{x\in\mathcal{X}_{\mathrm{test}}} \mathbf{1}\{a_{12}(x) = \arg\max_a u(x, a)\}$, using the same deterministic tie rule as the decision audit. Table 8 reports unweighted means and one standard deviation over five matched runs. Its decision-change column uses $|D|/|\mathcal{X}_{\mathrm{test}}|$, while Table 1 reports the consensus-conditional rate $|D|/|C|$. The utility-loss fraction column is h_D, and the final column is h_{all}. Threshold utility-loss intervals use 10,000 percentile bootstrap replicates clustered by chronological query index, resampling together all records that share an index across matched runs. Paired operating-point intervals use 5,000 bootstrap resamples of the five matched run pairs.

C.6 Experiment artifacts

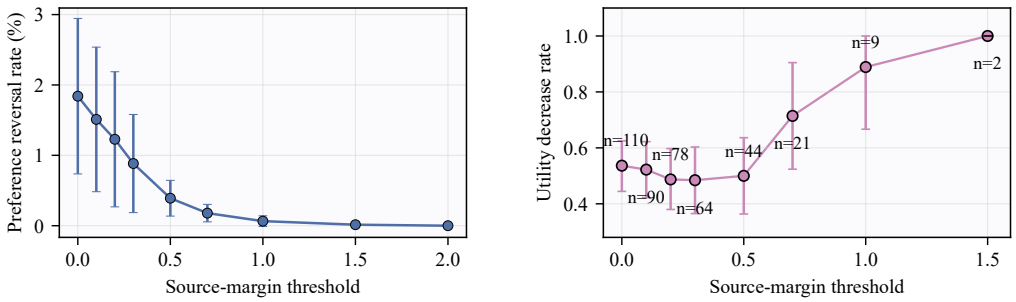
The experiment archive contains run-specific returns, fixed splits, representation initializations, per-run and summary outputs, plotting inputs, a result manifest, and a deterministic reproduction script. The result manifest records the source files, aggregation rules, statistical units, and configurations for the reported quantities. Tables 4–5 summarize the experiment roles and protocol settings.

D Additional Preference Reversal Analysis

D.1 Representative decision change from a preference reversal

One audited test query illustrates the link from a pairwise reversal to an action change. Both source readouts select $s = a_s = 0$, while the pooled readout selects $p = a_{12} = 1$. For the pair oriented toward the source-consensus action, the margins are $\Delta^1_{s,p} = +1.800$, $\Delta^2_{s,p} = +1.730$, and $\Delta^{12}_{s,p} = -0.086$. The pooled choice therefore changes a preference supported by both sources. Its negative margin also places this record outside the numerical tie band.

D.2 Source-margin sensitivity



(a) Strict reversal rate

(b) Utility-loss fraction

Figure 4: Strict-reversal rates and utility-loss fractions as the source-margin threshold increases. Panel (a) shows five-run variation. Panel (b) uses bootstrap intervals clustered by chronological query index. Higher thresholds retain fewer changed decisions, so the high-margin tail should be read together with its denominator.

Table 6: Changed decisions and utility outcomes by source-margin threshold. Clusters are distinct chronological query indices. Intervals use the cluster bootstrap. The highest thresholds retain few changed decisions.

τ	Changed	Clusters	Loss	Gain	h_D	Bootstrap 95% CI
0.0	110	98	59	51	0.536	[0.444, 0.625]
0.1	90	85	47	43	0.522	[0.422, 0.622]
0.2	78	75	38	40	0.487	[0.380, 0.597]
0.3	64	62	31	33	0.484	[0.365, 0.603]
0.5	44	44	22	22	0.500	[0.364, 0.636]
0.7	21	21	15	6	0.714	[0.524, 0.905]
1.0	9	9	8	1	0.889	[0.667, 1.000]
1.5	2	2	2	0	1.000	[1.000, 1.000]
2.0	0	0	0	0	–	–

Records with the same chronological test index across matched runs form one bootstrap cluster. Figure 4 and Table 6 show that reversals can persist after weak source preferences are excluded. The utility-loss fraction at the largest thresholds is less informative: at $\tau = 1.5$, only two loss records in two clusters remain, making its interval degenerate. The threshold sweep thus supports the presence of stronger-margin reversals without establishing a precise high-threshold loss rate.

D.3 Controlled reversal and tie counts

Table 7 lists selected settings from the main controlled sweep. Eligible pairs, strict reversals, and pooled ties are summed over five replications, while Gram mismatch and strict preference-preservation rate are unweighted means of the five replication-level values used in Figure 2. Because these summaries use different aggregation levels, dividing the pooled failure count by the pooled eligible count gives a different quantity from the reported mean strict preference-preservation rate.

Table 7: Selected settings from the controlled main sweep. Counts are pooled over replications, while Gram mismatch and strict preference-preservation rate are means of replication-level values. Strict preference reversals account for nearly all failures. Pooled ties are rare.

H	N	α	Eligible	Mean δ_{norm}	Mean preservation	Reversals	Ties
2	600	0.00	10,712	0.072	1.000	0	0
2	600	0.05	10,996	0.498	1.000	0	0
2	600	0.10	9,635	0.840	0.949	517	1
2	600	0.15	9,369	1.070	0.073	8,632	0
2	600	0.20	10,638	1.217	0.240	7,991	0
2	600	0.25	10,601	1.295	0.231	8,072	1
2	600	0.30	8,696	1.340	0.063	7,992	0
2	600	0.50	10,088	1.405	0.145	8,386	0
2	600	0.90	8,785	1.414	0.000	8,785	0
2	600	1.00	10,391	1.414	0.111	9,051	7

At low Gram mismatch, pooled refitting almost always preserves the shared source preference. Beyond the sampled transition, strict reversals account for nearly all failures. The later non-monotone values show that the first crossing is specific to the evaluated grid and configuration. Gram mismatch is rounded to three decimals. Entries printed as 1.414 remain strictly below $\sqrt{2}$ before rounding.

E Additional Geometry Regulation Results

E.1 Complete geometry-regulation configurations

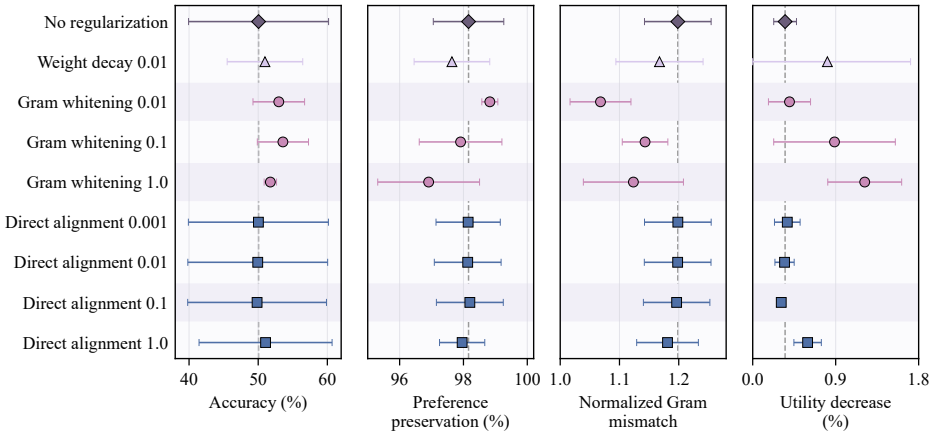


Figure 5: Five-run means and one-standard-deviation intervals for accuracy, strict preference-preservation rate, normalized Gram mismatch, and overall utility-loss rate across all nine configurations. No configuration gives the best point estimate on all four metrics.

Table 8: All training configurations under matched load returns (mean ± one standard deviation over five runs). Panel A reports accuracy, preservation, and geometry. Panel B reports decision changes and utility. Decision change uses $|D|/|\mathcal{X}_{\text{test}}|$.

Panel A: prediction and Gram geometry

Objective	λ_{reg}	Acc. (%)	Preservation (%)	δ_{norm}
No regularization	0	50.05±10.13	98.16±1.10	1.1984±0.0562
Weight decay	0.01	50.98±5.47	97.64±1.19	1.1675±0.0736
Gram whitening	0.01	52.98±3.73	98.83±0.25	1.0680±0.0514
Gram whitening	0.1	53.57±3.70	97.91±1.30	1.1432±0.0385
Gram whitening	1.0	51.76±0.90	96.91±1.60	1.1235±0.0844
Direct alignment	0.001	50.03±10.13	98.15±1.01	1.1985±0.0563
Direct alignment	0.01	49.93±10.11	98.13±1.05	1.1982±0.0563
Direct alignment	0.1	49.84±10.04	98.20±1.05	1.1964±0.0561
Direct alignment	1.0	51.07±9.62	97.96±0.71	1.1811±0.0521

Panel B: decision and utility outcomes

Objective	λ_{reg}	Decision change (%)	h_D (%)	h_{all} (%)
No regularization	0	0.66±0.25	55.19±6.96	0.35±0.12
Weight decay	0.01	1.48±1.48	55.38±16.08	0.81±0.90
Gram whitening	0.01	0.74±0.31	50.47±13.30	0.40±0.23
Gram whitening	0.1	1.63±1.33	56.91±10.65	0.89±0.66
Gram whitening	1.0	1.98±0.39	60.12±10.91	1.22±0.40
Direct alignment	0.001	0.71±0.28	53.32±7.53	0.38±0.14
Direct alignment	0.01	0.66±0.22	54.55±10.48	0.35±0.10
Direct alignment	0.1	0.67±0.15	47.62±9.20	0.31±0.05
Direct alignment	1.0	1.08±0.33	55.93±4.77	0.60±0.15

Figure 5 and Table 8 place all nine configurations on common accuracy, preservation, geometry, and utility scales. The configuration that best preserves pairwise preferences does not minimize observed utility loss: whitening at $\lambda_{\text{reg}} = 0.01$ has the highest strict preference-preservation rate and lowest Gram mismatch, whereas direct alignment at 0.1 has the lowest utility-loss point estimate. The distinction matters because a changed preference need not reduce task utility. The table's decision-

change rate uses $|D|/|\mathcal{X}_{\text{test}}|$. The consensus-conditional rate $|D|/|C|$ is reported in Table 1. Paired comparisons with no regularization are given in Table 9.

E.2 Paired uncertainty estimates

Table 9: Paired five-run differences from the no-regularization baseline for overall utility-loss rate and normalized Gram mismatch. Several mismatch intervals exclude zero. Most utility-loss intervals at moderate strengths include zero.

Method	λ_{reg}	Mean diff.	Paired t 95% CI	Bootstrap 95% CI
Panel A: Δh_{all}, percentage points				
Direct alignment	0.001	+0.024	[-0.047, +0.095]	[-0.012, +0.071]
Direct alignment	0.01	-0.006	[-0.061, +0.049]	[-0.036, +0.030]
Direct alignment	0.1	-0.042	[-0.166, +0.083]	[-0.125, +0.036]
Direct alignment	1.0	**+0.244**	**[+0.118, +0.371]**	**[+0.173, +0.316]**
Gram whitening	0.01	+0.048	[-0.144, +0.239]	[-0.066, +0.179]
Gram whitening	0.1	+0.536	[-0.156, +1.228]	[+0.107, +0.965]
Gram whitening	1.0	**+0.864**	**[+0.405, +1.322]**	**[+0.530, +1.126]**
Weight decay	0.01	+0.459	[-0.799, +1.716]	[-0.048, +1.370]
Panel B: $\Delta \delta_{\text{norm}}$				
Direct alignment	0.001	+0.0001	[-0.0001, +0.0002]	[-0.0000, +0.0001]
Direct alignment	0.01	-0.0003	[-0.0007, +0.0002]	[-0.0006, -0.0000]
Direct alignment	0.1	-0.0020	[-0.0047, +0.0007]	[-0.0039, -0.0008]
Direct alignment	1.0	**-0.0174**	**[-0.0316, -0.0031]**	**[-0.0269, -0.0097]**
Gram whitening	0.01	**-0.1304**	**[-0.2560, -0.0049]**	**[-0.1884, -0.0396]**
Gram whitening	0.1	-0.0553	[-0.1611, +0.0506]	[-0.1059, +0.0201]
Gram whitening	1.0	-0.0749	[-0.2219, +0.0720]	[-0.1679, +0.0189]
Weight decay	0.01	**-0.0309**	**[-0.0613, -0.0005]**	**[-0.0495, -0.0127]**

Table 9 compares each configuration with no regularization within the same run. Several geometry changes remain visible under both interval procedures, but most utility-loss intervals at moderate strengths include zero. The paired analysis therefore supports a clearer effect on measured geometry than on the downstream loss rate. At the largest whitening and direct-alignment strengths, both procedures indicate increased utility loss, limiting how aggressively either penalty should be applied.

F Additional Cross-Domain Analysis

F.1 Complete cross-domain compatibility results

Table 10: Five-run cross-domain preference-preservation audit. Eligible-pair and reversal counts are pooled descriptive totals. Gram mismatch and strict preference-preservation rate are summarized across runs. Similar mismatch can accompany markedly different realized reversal rates.

Domain	Partition	Eligible	Reversals	δ_{norm}	Preservation
Default Credit	Age	4,447	0	0.099±0.035	1.0000±0.0000
Default Credit	Education	4,442	1	0.160±0.030	0.9998±0.0005
Default Credit	Recent delinquency	4,268	4	1.036±0.211	0.9991±0.0015
Default Credit	Recorded sex	4,434	2	0.193±0.068	0.9995±0.0006
Heart Disease	Age	577	40	0.842±0.066	0.9306±0.0188
Heart Disease	Recorded sex	518	55	0.903±0.040	0.8922±0.0458

Table 10 gives the six partitions plotted in Figure 3. Pooled ties are absent in all six. The four Default Credit partitions remain close to complete preservation, whereas both Heart Disease partitions contain strict reversals. The relationship is informative across these settings but does not define a dataset-independent mismatch threshold: the same distance can have different consequences when source evidence and query margins differ. Eligible-pair and reversal counts are pooled descriptive totals. Gram mismatch and strict preference-preservation rate use the five experimental runs as replication units.