DO FUNCTIONALLY EQUIVALENT REPRESENTATIONS FAIL ALIKE? TASK-CONSISTENT GAUGES UNDER PARTIAL AVAILABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Intermediate representations are often only partly available in split, adaptive, or progressive inference. If two coordinate systems implement exactly the same complete predictor, should they also fail alike when some representation blocks are missing? We show that they need not: function-preserving rotations can leave every complete prediction unchanged yet produce sharply different partial predictions. Under fixed-cardinality availability, partial logits and pairwise class margins are unbiased estimates of their complete counterparts. Their variance separates how much representation remains from how task-relevant evidence is distributed across coordinates. This explains why reducing global logit-estimation variance can still hurt classification. The nearby margins that determine the decision may become less stable. We therefore introduce BoundaryGauge, which preserves the complete predictor while selecting coordinates that stabilize those margins, without ground-truth labels or training on partial masks. At 50% availability, the same fixed Hadamard gauge changes accuracy by -13.5 points for one ResNet50 predictor but $+24.9$ points for another. BoundaryGauge adapts successfully to both. Across four architectures and 48 frozen-predictor cases, each optimized gauge improves mean, P05, and exact worst-subset accuracy across all three tested availability levels. Partial robustness therefore depends not only on the complete function, but on how its decision geometry is distributed across coordinates.

1 INTRODUCTION

Modern inference systems often operate with incomplete intermediate representations. In split inference, activations may be transmitted across devices; adaptive pipelines may keep only selected components; and progressive transmission may expose only part of a representation at a given time. For example, over a bandwidth-constrained or unreliable edge link, only a subset of activation blocks may be available in time for downstream computation (Itahara et al., 2021). Likewise, a progressive pipeline may begin downstream inference before the complete representation is available (Lan et al., 2023). These settings raise a basic question: *if only part of a hidden representation is available, does the choice of representation coordinates matter?*

For the complete predictor, the usual answer is no. Hidden features can be permuted or rotated while the following layer is adjusted to compensate, leaving the realized function unchanged (Lenc & Vedaldi, 2015; Kornblith et al., 2019; Bansal et al., 2021; Entezari et al., 2022). Function-preserving rotations are already useful in settings such as quantization, where the same model can become easier to compress in one coordinate system than another (Ashkboos et al., 2024; Liu et al., 2025). With the full representation present, these coordinate systems are exactly interchangeable.

Partial availability changes the picture. This matters whenever downstream decisions must be made before the complete representation is available. Figure 1 shows the key idea. A rotation and its compensated head cancel when every block is present, but removing blocks between them breaks that cancellation. Two representations can therefore implement the same complete predictor and still produce different decisions under the same missing-block pattern. Partial availability turns an otherwise invisible coordinate freedom into observable differences in task behavior. Under partial

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.

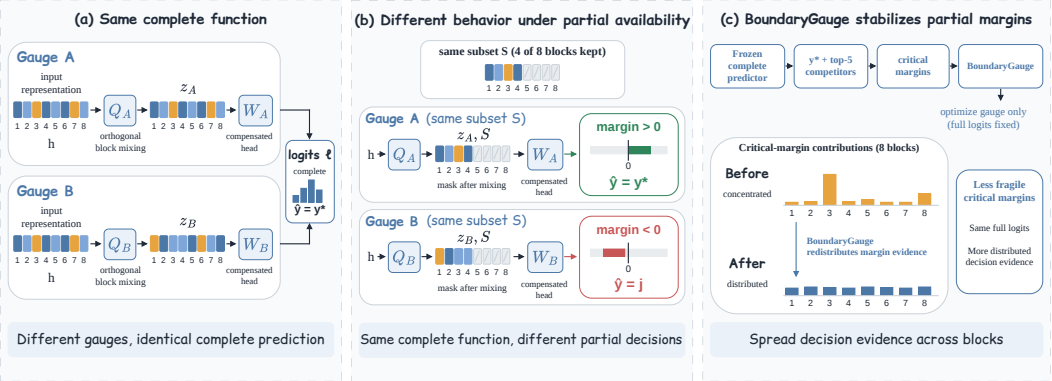

Figure 1: **Full-function equivalence does not determine partial behavior.** A block-axis gauge and compensated head leave complete logits unchanged (left). Removing the same blocks can expose different partial decisions (center). BoundaryGauge chooses an equivalent coordinate system that spreads decision-critical evidence more evenly across blocks (right).

availability, choosing among functionally equivalent representations becomes a meaningful design degree of freedom.

Existing robustness methods typically change training, transmission, reconstruction, or feature selection to cope with missing features (Srivastava et al., 2014; Morvan et al., 2021; Lan et al., 2023; Itahara et al., 2021; Yang et al., 2025). We ask a different question: can partial robustness improve while the complete predictor remains exactly fixed, solely by choosing among coordinate systems that implement the same function?

The answer becomes clear when the complete output is written as a sum of block contributions. A partial predictor observes only a subset of those contributions and rescales them to estimate the full output. Changing coordinates cannot change the complete sum, but it can change how strongly each block contributes. What matters is not only how much of the representation remains, but where the predictor's decision evidence is distributed across blocks.

For classification, the critical quantity is the margin separating the predicted class from its strongest competitors. Reducing overall logit error does not necessarily protect these margins: the total error can become smaller while a nearby decision boundary becomes easier to cross. This gives us a task-specific target: stabilize the margins that can actually change the prediction.

We build on this idea with BoundaryGauge. For each calibration example, the frozen predictor identifies its leading competing classes. BoundaryGauge then chooses a function-preserving rotation that distributes the corresponding margin contributions more evenly across representation blocks. The backbone and task head remain frozen, and analytical compensation preserves the complete logits exactly. Because the objective does not depend on a particular retained fraction, one learned gauge can be used across multiple availability levels.

The experiments support this account from three directions. First, there is no universally robust generic rotation: at 50% availability, the same Hadamard gauge changes accuracy by -13.5 points for one ResNet50 predictor and $+24.9$ points for another. Second, minimizing global logit variance can improve that surrogate while making actual decisions less stable, whereas margin-aligned optimization improves both decision fidelity and task accuracy. Third, across four architectures and 48 frozen-predictor settings, a single learned gauge per predictor improves mean, 5th-percentile, and exact worst-subset accuracy at 25%, 50%, and 75% availability. On the hard native ResNet50 case, it also approaches the performance of a supervised mask-trained control without using labels or partial masks during gauge optimization.

Contributions. This work makes three main contributions. (1) We show that representations that are exactly equivalent for complete inference can fail differently under partial availability, and formalize partial prediction as estimation from block contributions. (2) We identify class-margin stability as the task-relevant geometry for classification, explaining why a global logit-variance objective can

move in the wrong direction. (3) We introduce BoundaryGauge, a function-preserving coordinate optimization method that uses the frozen predictor's own decision structure and generalizes across retained fractions and architectures without partial-mask training.

2 Equivalent Coordinates Under Partial Availability

In this section, we define the function-preserving coordinate family and show why partial availability makes its members behave differently.

Complete predictor. We begin with representations that differ internally but implement the same complete predictor. Divide the final representation into B equal blocks, $h = [h_1; \ldots; h_B]$, with $h_i \in \mathbb{R}^r$, followed by a linear task head for C classes

$$\ell(h) = Wh + b, \qquad W \in \mathbb{R}^{C \times Br}. \tag{1}$$

For an orthogonal $Q \in \mathbb{R}^{B \times B}$, let I_r denote the $r \times r$ identity and define $T_Q = Q \otimes I_r$, $z = T_Q h$, and the compensated head $W_Q = WT_Q^\top$. Then

$$W_Q z + b = WT_Q^\top T_Q h + b = Wh + b. \tag{2}$$

Thus Q changes the internal coordinates without changing the complete logits. We call the resulting family *function-preserving representation gauges*. Appendix D verifies the invariance numerically.

Partial predictor. The gauge becomes visible only after some blocks are unavailable. Let $[B] = \{1, \ldots, B\}$. For $S \subseteq [B]$ with $|S| = k$, let P_S retain blocks in S and zero the rest. We study

$$\hat{\ell}_{Q,S} = b + \frac{B}{k} W_Q P_S T_Q h. \tag{3}$$

The B/k factor rescales the retained contribution to estimate the full block sum. With all blocks present, Eq. (3) reduces to the same complete predictor for every gauge. With $k < B$, however, the masking operator sits between the coordinate transform and its compensation, so $W_{Q_1} P_S T_{Q_1}$ and $W_{Q_2} P_S T_{Q_2}$ generally differ. Exact full-function equivalence therefore does not imply equal partial predictors.

We study the setting in which downstream computation receives a fixed-cardinality subset of representation blocks. The experiments isolate this representation-level effect; systems-level compute savings depend on how those blocks are produced and transmitted.

3 Partial Prediction as Estimation

In this section, we show that partial prediction has a simple sampling interpretation. Write the complete logits as contributions from individual blocks. For a fixed gauge, let $c_i = c_i(Q)$ denote the contribution of block i:

$$Wh = \sum_{i=1}^{B} c_i, \qquad c_i \in \mathbb{R}^C. \tag{4}$$

Then Eq. (3) is

$$\hat{\ell}_S = b + \frac{B}{k} \sum_{i \in S} c_i. \tag{5}$$

Equation (5) is sampling without replacement from the B block contributions (Serfling, 1974). A gauge leaves their total unchanged but can redistribute that total across blocks.

Proposition 1 (Unbiased partial logits)**.** *If S is uniform over all k-subsets of $[B]$, then $\mathbb{E}_S[\hat{\ell}_S] = \ell$.*

Proof sketch. Every block is included with probability k/B, so $\mathbb{E}[(B/k) \sum_{i \in S} c_i] = \sum_i c_i$; b is deterministic. Complete proofs appear in Appendix A. □

Proposition 2 (Exact fixed-cardinality variance)**.** *For any $u \in \mathbb{R}^C$, define the scalar projection $a_i = u^\top c_i$ with $\bar{a} = B^{-1}\sum_i a_i$,*

$$\mathrm{Var}_S\left[\frac{B}{k}\sum_{i\in S} a_i\right] = \alpha(B,k)\sum_{i=1}^{B}(a_i-\bar{a})^2, \quad \alpha(B,k) = \frac{B(B-k)}{k(B-1)}. \tag{6}$$

Equation (6) separates two factors: how much of the representation remains and how unevenly the relevant signal is distributed across blocks. The retained cardinality appears only through the positive scalar $\alpha(B,k)$. For fixed task directions, the gauge that minimizes this variance therefore does not depend on k.

3.1 Classification depends on class margins

A first idea is to spread every logit coordinate as evenly as possible across blocks. Applying Eq. (6) to all logit coordinates gives the natural global objective

$$J_{\mathrm{logit}}(Q) = \mathbb{E}_x \sum_i \|c_i(Q) - \bar{c}(Q)\|_2^2. \tag{7}$$

where $\bar{c}(Q) = B^{-1}\sum_i c_i(Q)$. This objective exactly minimizes expected squared logit-estimation error. But a classifier does not fail because the logit vector moves in an arbitrary Euclidean direction; it fails when that movement crosses a class boundary. Global error can therefore shrink while the few directions closest to changing the decision become less stable.

Let $y^\star = \arg\max_c \ell_c$ and let $j \neq y^\star$ be a competitor. Define the full margin and blockwise margin contribution

$$m_j = \ell_{y^\star} - \ell_j, \qquad a_{i,j} = c_{i,y^\star} - c_{i,j}, \qquad \bar{a}_j = B^{-1}\sum_i a_{i,j}. \tag{8}$$

The partial margin is

$$\hat{m}_{j,S} = b_{y^\star} - b_j + \frac{B}{k}\sum_{i\in S} a_{i,j}. \tag{9}$$

Proposition 3 (Partial decision margins)**.** *For every competitor j,*

$$\mathbb{E}_S[\hat{m}_{j,S}] = m_j, \quad \mathrm{Var}_S[\hat{m}_{j,S}] = \alpha(B,k)\sum_i (a_{i,j}-\bar{a}_j)^2. \tag{10}$$

A class change occurs when at least one pairwise margin crosses zero. Equation (10) therefore focuses the analysis on the directions that can change the prediction. In Section 6, we test this link directly using exact subset-induced decision flips.

4 BoundaryGauge

In this section, we turn the margin analysis into a coordinate-selection method. BoundaryGauge uses the frozen predictor to identify the class boundaries most relevant to each calibration example and then chooses coordinates that make those margins less sensitive to missing blocks. For each calibration sample x, we compute the invariant complete prediction $y^\star(x)$ and the M highest-scoring competitors $\mathcal{J}_M(x)$. Because complete logits are invariant to Q, these target classes remain fixed throughout gauge optimization. We minimize

$$J_{\mathrm{BG}}(Q) = \mathbb{E}_x \frac{1}{M}\sum_{j\in\mathcal{J}_M(x)}\sum_{i=1}^{B}\left(a_{i,j}(Q) - \bar{a}_j(Q)\right)^2. \tag{11}$$

Here $\bar{a}_j(Q) = B^{-1}\sum_i a_{i,j}(Q)$. By Eq. (10), this objective reduces the average variance of the selected partial margins. It contains no retained cardinality k, so the same optimized gauge applies across availability levels. Our primary instance uses $M = 5$ and is denoted BoundaryGauge-5; Appendix C.5 reports diagnostics for other competitor counts.

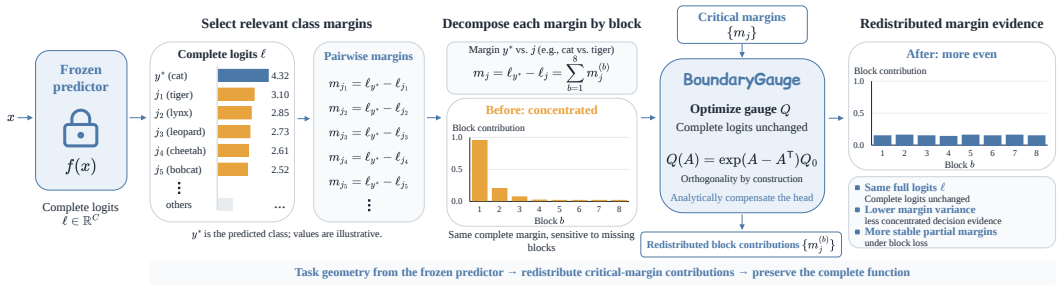

Figure 2: **BoundaryGauge mechanism.** The frozen predictor identifies critical margins; BoundaryGauge redistributes their block contributions while analytical head compensation preserves the complete logits.

We parameterize $Q(A) = \exp(A - A^\top)Q_0$ with $A \in \mathbb{R}^{B\times B}$, which keeps the transform orthogonal throughout optimization. Only A changes. The backbone and task head remain frozen, while the head is analytically compensated for the current gauge. Optimization uses complete representations and the frozen predictor's own top classes; it does not use ground-truth labels, sampled availability masks, or masked forward passes. For comparison, MASKEDCE uses the same gauge parameterization but is trained with labels and sampled partial masks.

5 EXPERIMENTS

In this section, we evaluate whether coordinate choice changes partial robustness, whether margin alignment explains the effect, and whether one learned gauge transfers across availability levels and architectures.

Predictor regimes. We evaluate two complementary regimes while keeping the backbone and task head frozen during gauge optimization. **Frozen probes** use ImageNet-pretrained ResNet50 (He et al., 2016), ConvNeXt-Tiny (Liu et al., 2022), ViT-B/16 (Dosovitskiy et al., 2021), and Swin-T (Liu et al., 2021) as frozen feature extractors, with linear CIFAR-100 (Krizhevsky, 2009) probes trained only on complete representations in three independent runs. **Native predictors** retain the original ImageNet-1K (Deng et al., 2009) classifier heads and official preprocessing, and evaluate on ImageNetV2 MatchedFrequency (Recht et al., 2019). The native data are deterministically split into equally sized calibration and held-out evaluation sets. Appendix B gives the full preprocessing, optimization, block-count, and numerical-audit details.

Availability and gauges. The main experiment fixes $B = 8$ and exhaustively evaluates all $\binom{8}{2} = 28$, $\binom{8}{4} = 70$, and $\binom{8}{6} = 28$ subsets, corresponding to 25%, 50%, and 75% availability. We compare Identity, a fixed Hadamard transform, four precommitted random orthogonal gauges, GLOBALVAR (Eq. 7), BoundaryGauge-5, and MASKEDCE. Paired learned methods start from the same stored random matrices and are optimized on their calibration objectives. Evaluation is performed only after the gauge is fixed.

Metrics. We use complementary metrics so that average performance does not hide fragile subsets:

- **Mean accuracy:** task accuracy averaged over all subsets at a fixed cardinality.
- **5th-percentile accuracy:** the 5th percentile across subsets, capturing lower-tail robustness.
- **Worst-subset accuracy:** the minimum accuracy over the exhaustively enumerated subsets.
- **Fidelity and decision flips:** fidelity is agreement with the complete predictor, and flip probability is one minus fidelity.

For cross-cardinality analysis, we report whether the same final Q improves a metric over its paired random gauge at all three availability levels.

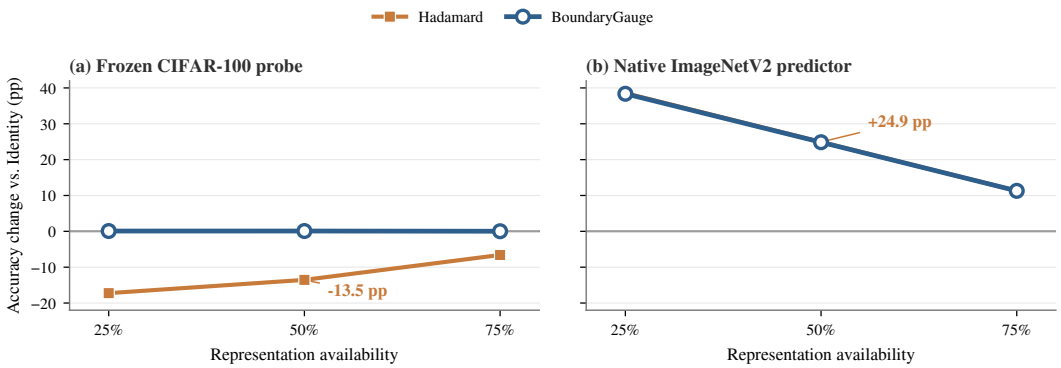

Figure 3: **The same generic gauge reverses across predictors.** Accuracy change relative to Identity on the same ResNet50 architecture. Hadamard is harmful for the frozen CIFAR-100 predictor but beneficial for the native ImageNetV2 predictor. BoundaryGauge-5 stays near Identity in the former and tracks the useful mixed basis in the latter.

6 RESULTS

In this section, we follow the paper's central causal chain: generic coordinate mixing is task dependent, global logit variance can optimize the wrong quantity, and margin-aligned coordinate selection repairs the resulting failures.

6.1 A FIXED GENERIC GAUGE CAN HELP ONE PREDICTOR AND HURT ANOTHER

Figure 3 gives the clearest evidence that no fixed mixing rule is universally robust. The architecture and partial-availability operator are unchanged, yet Hadamard is strongly harmful for the frozen CIFAR-100 probe and strongly beneficial for the native ImageNetV2 head. At 50% availability, the change relative to Identity is -13.5 versus $+24.9$ percentage points. The difference comes from how each predictor's task evidence is distributed in the chosen basis. BoundaryGauge-5 adapts to both regimes: it stays near Identity when mixing is harmful and moves toward the useful mixed solution when mixing helps.

6.2 GLOBAL LOGIT VARIANCE CAN MOVE IN THE WRONG DIRECTION

The native ResNet50 explains why this task dependence matters. GLOBALVAR lowers the global centered-logit variance from every matched random initialization, yet nearest-competitor margin variance increases. At 50% availability, all four initializations consequently show more decision flips and lower accuracy. The complete predictor can therefore have smaller overall logit error while becoming less stable in the class-separation directions that determine the final decision.

Figure 4 makes the distinction visible. Across the 14 native ResNet50 gauges at 50% availability, global centered-logit variance is poorly aligned with exact decision flips (Spearman $\rho = -0.20$), whereas nearest-competitor margin variance is perfectly rank aligned ($\rho = 1.00$). The paired interventions then show the task consequence directly: BoundaryGauge-5 and MASKEDCE reduce flips and raise partial accuracy, while GLOBALVAR moves in the opposite direction.

6.3 BOUNDARYGAUGE REPAIRS THE HARD NATIVE CASE WITHOUT MASK TRAINING

Figure 5 shows the resulting repair. BoundaryGauge-5 improves mean partial accuracy at all three availability levels and nearly overlaps the supervised MASKEDCE reference. At 50% availability, it gains about 6.6 points over the paired random gauges and finishes within 0.1 point of MASKEDCE. Its exact flip probability drops from 24.2% to 10.5%, while GLOBALVAR increases flips to 27.8%. The corresponding worst-subset results are reported in Appendix C.

The same pattern holds across all starts and retained fractions: BoundaryGauge-5 improves task accuracy and fidelity to the complete predictor in all 12 combinations of random initialization and

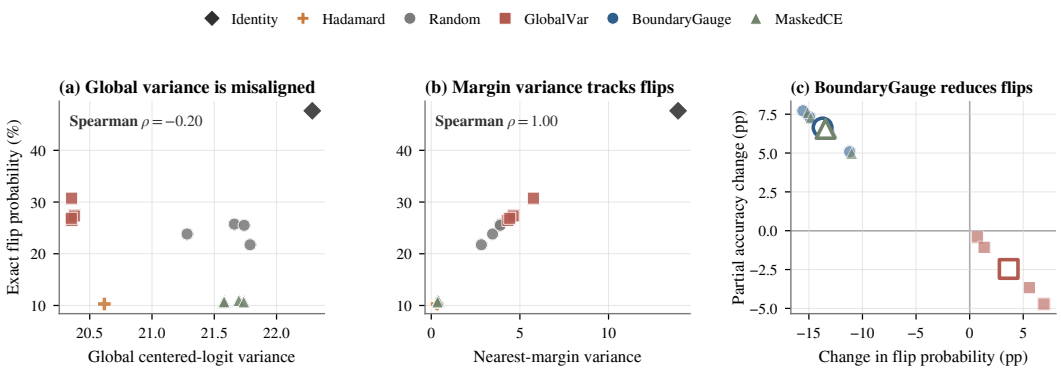

Figure 4: **Decision-margin geometry identifies the relevant failure signal.** At 50% availability on native ResNet50, (a) global centered-logit variance is weakly related to exact decision flips across the 14 evaluated gauges, while (b) nearest-competitor margin variance is perfectly rank aligned with them. (c) Paired changes from the four random initializations show the resulting task effect: BoundaryGauge-5 and supervised MASKEDCE reduce flips and improve accuracy, whereas GLOBALVAR increases flips and lowers accuracy. Hollow markers denote method means in panel (c).

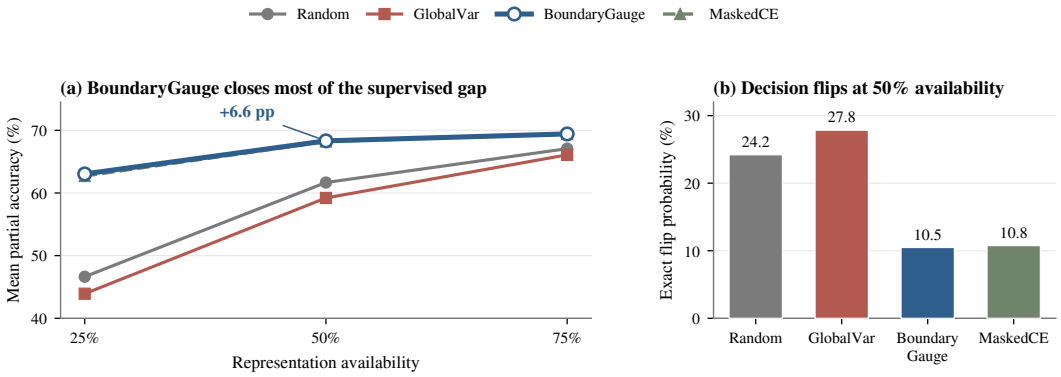

Figure 5: **BoundaryGauge repairs the hard native ResNet50 case.** (a) Mean ImageNetV2 partial accuracy across availability levels, averaged over four matched initializations for learned and random gauges. BoundaryGauge-5 closes most of the gap to supervised MASKEDCE without partial-mask training. (b) At 50% availability, the gain coincides with a large reduction in exact decision flips.

availability level for the native ResNet50. The frozen complete predictor is therefore sufficient to identify useful class-separation directions for gauge optimization, even without training on partial masks.

6.4 ONE GAUGE GENERALIZES ACROSS AVAILABILITY AND ARCHITECTURES

The frozen-predictor regime asks whether each availability level needs its own coordinate system. Across all four architectures, the same final gauge improves mean accuracy at 25%, 50%, and 75% availability simultaneously in all **48** combinations of architecture, probe run, and random initialization. All 48 combinations also improve 5th-percentile and exact worst-subset accuracy at every availability level. Figure 6 shows that gains are largest where random gauges leave substantial headroom, but remain positive across all architectures and all three cardinalities.

This behavior matches Eq. (10): changing k rescales the variance of a fixed task direction without changing the contribution pattern that the gauge organizes. Native ConvNeXt-Tiny, ViT-B/16, and Swin-T provide the complementary low-headroom case. When the original predictor is already insensitive to gauge choice, BoundaryGauge-5 remains nearly neutral. BoundaryGauge therefore produces its largest gains when critical decision evidence is concentrated across blocks and changes little when that evidence is already well distributed.

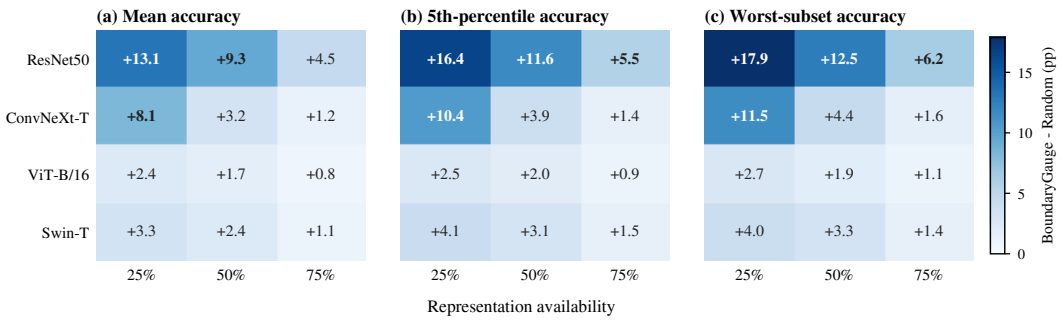

Figure 6: **One learned gauge transfers across availability levels and robustness metrics.** Gain of BoundaryGauge-5 over the paired random gauge, averaged over three probe runs and four random initializations, for mean, 5th-percentile, and exact worst-subset accuracy. The same final gauge is evaluated at 25%, 50%, and 75% availability. Gains remain positive across all four architectures and all three metrics.

7 Related Work

In this section, we position BoundaryGauge relative to work on representation equivalence, function-preserving reparameterization, and robustness to missing intermediate features.

Representation similarity and functional equivalence. Neural representations are commonly compared through alignment or similarity measures such as canonical correlation analysis (CCA), centered kernel alignment (CKA), and generalized shape metrics (Lenc & Vedaldi, 2015; Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019; Williams et al., 2021). Other work evaluates representations through functional behavior, learned stitching, or shared invariances (Ding et al., 2021; Klabunde et al., 2025; Bansal et al., 2021; Hernandez et al., 2023; Nanda et al., 2022). Our setting starts from exact functional equivalence under a known analytical compensation and studies what happens after a partial-availability operator is inserted between the representation and its head.

Neural symmetry and function-preserving reparameterization. Permutation and related parameter symmetries create many internal parameterizations of the same function and have been used for alignment, mode connectivity, and model merging (Entezari et al., 2022; Ainsworth et al., 2023; Jordan et al., 2023; Godfrey et al., 2022). Recent work further shows that hidden-state measurements can vary under function-preserving changes of basis (Cain, 2026; Karasan, 2026). BoundaryGauge uses the same freedom for a different purpose: selecting an equivalent basis whose block contributions are more stable when some blocks are unavailable.

Optimizing function-preserving coordinates. QuaRot and SpinQuant rotate Transformer representations to reduce quantization sensitivity, while Codec-Gauge chooses compression-friendly gauges for key-value caches (Ashkboos et al., 2024; Liu et al., 2025; Jiang et al., 2026). These methods share the principle that a transformation invisible to complete inference can matter once a coordinate-sensitive operator is introduced. Here that operator is block removal. Its sampling structure lets us connect coordinate choice directly to the variance of class margins. Canonicalization methods address a related symmetry problem by selecting representatives to enforce invariance or equivariance (Kaba et al., 2023; Tahmasebi & Jegelka, 2025; Li et al., 2025; Shumaylov et al., 2025; Lin & Levie, 2026); BoundaryGauge instead preserves the complete predictor and changes only its behavior under partial availability.

Missing and partially transmitted features. Dropout and missing-feature methods train models to tolerate deletion or infer unavailable inputs (Srivastava et al., 2014; Morvan et al., 2021). Split-inference systems similarly learn which features to transmit, simulate packet loss during training, or equalize semantic information across transmitted channels (Lan et al., 2023; Itahara et al., 2021; Yang et al., 2025). BoundaryGauge complements these approaches by acting entirely within a fixed complete predictor: it changes neither the backbone nor the task head, but selects an equivalent coordinate system whose decision evidence is better distributed across blocks.

8 Limitations and Conclusion

The current study focuses on blockwise orthogonal gauges before a linear classification head, uniform fixed-cardinality availability, and vision models. Extending the same idea to structured or non-uniform availability, nonlinear downstream computation, language and multimodal models, and independently producible transformed blocks is a natural next step. The experiments isolate representation-level robustness; end-to-end system gains will additionally depend on how partial representations are computed and communicated.

The main conclusion is straightforward: exact equivalence under complete inference does not determine behavior under partial availability. Missing blocks expose how a predictor's decision evidence is distributed across representation coordinates. For classification, organizing that evidence around the margins that separate competing classes is more useful than minimizing global logit error alone. BoundaryGauge turns this observation into a function-preserving coordinate optimization method that improves partial robustness without changing the complete predictor or training on partial masks.

AI Use Statement

The authors used large language model-based AI tools to assist with writing and language polishing, and with the retrieval and discovery of potentially relevant related work. The authors carried out the core scientific work, including problem formulation, method development, implementation, experimental design and execution, and the analysis and interpretation of results. All cited sources were checked against the original literature. The authors reviewed the final manuscript and take full responsibility for its content and claims.

Reproducibility Statement

The supplementary archive contains a clean implementation of the gauge transform, exact partial-logit and margin estimators, the BoundaryGauge-5 objective, synthetic identity tests, and scripts that reproduce the paper's tables and figures from the packaged structured evidence. Appendix A gives complete proofs of the estimator identities; Appendix B specifies data splits, preprocessing, optimization, gauges, and exact subset evaluation; Appendix D records the float32 and float64 function-invariance checks. The submission package excludes model weights and datasets but identifies all public datasets and pretrained model variants needed to reproduce the experiments.

References

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *ICLR*, 2023.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *NeurIPS*, 2024. doi: 10.52202/079017-3180.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *NeurIPS*, 2021.

Jericho Cain. Gauge freedom and metric dependence in neural representation spaces. *arXiv preprint arXiv:2603.06774*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *NeurIPS*, volume 34, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *ICLR*, 2022.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. In *NeurIPS*, volume 35, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Adriano Hernandez, Rumen Dangovski, Peter Y. Lu, and Marin Soljacic. Model stitching: Looking for functional similarity between representations. *arXiv preprint arXiv:2303.11277*, 2023.

Sohei Itahara, Takayuki Nishio, and Koji Yamamoto. Packet-loss-tolerant split inference for delay-sensitive deep learning in lossy wireless networks. In *IEEE GLOBECOM*, 2021. doi: 10.1109/GLOBECOM46510.2021.9685179.

Yitao Jiang, Yaoqing Yang, Luyang Zhao, Muhao Chen, and Devin J. Balkcom. Codec-Gauge: Learning compression-friendly gauges for Transformer KV caches. *arXiv preprint arXiv:2607.20538*, 2026.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: REnormalizing permuted activations for interpolation repair. In *ICLR*, 2023.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *ICML*, pp. 15546–15566, 2023.

Abdullah Karasan. Representation measurements under function-preserving reparameterizations. *arXiv preprint arXiv:2608.27020*, 2026.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. doi: 10.1145/3728458.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, pp. 3519–3529, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Qiao Lan, Qunsong Zeng, Petar Popovski, Deniz Gündüz, and Kaibin Huang. Progressive feature transmission for split classification at the wireless edge. *IEEE Transactions on Wireless Communications*, 22(6):3837–3852, 2023. doi: 10.1109/TWC.2022.3221778.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *CVPR*, pp. 991–999, 2015.

Yikang Li, Yeqing Qiu, Yuxuan Chen, and Zhouchen Lin. Affine steerable equivariant layer for canonicalization of neural networks. In *ICLR*, 2025.

Ya-Wei Eileen Lin and Ron Levie. Adaptive canonicalization with application to invariant anisotropic geometric networks. In *ICLR*, 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *ICLR*, 2025.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, pp. 11976–11986, 2022.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *NeurIPS*, pp. 5732–5741, 2018.

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What's a good imputation to predict with missing values? In *NeurIPS*, volume 34, pp. 11530–11540, 2021.

Vedant Nanda, Till Speicher, Camila Kolling, John P. Dickerson, Krishna Gummadi, and Adrian Weller. Measuring representational robustness of neural networks through shared invariances. In *ICML*, pp. 16368–16382, 2022.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *NeurIPS*, pp. 6076–6085, 2017.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, pp. 5389–5400, 2019.

Robert J. Serfling. Probability inequalities for the sum in sampling without replacement. *Annals of Statistics*, 2(1):39–48, 1974.

Zakhar Shumaylov, Peter Zaika, James Rowbottom, Ferdia Sherry, Melanie Weber, and Carola-Bibiane Schönlieb. Lie algebra canonicalization: Equivariant neural operators under arbitrary Lie groups. In *ICLR*, 2025.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(56):1929–1958, 2014.

Behrooz Tahmasebi and Stefanie Jegelka. Generalization bounds for canonicalization: A comparative study with group averaging. In *ICLR*, 2025.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *NeurIPS*, volume 34, pp. 4738–4750, 2021.

Xiao Yang, Shuai Ma, Yong Liang, and Guangming Shi. Feature partitioning and semantic equalization for intrinsic robustness in semantic communication under packet loss. *arXiv preprint arXiv:2511.16983*, 2025.

A Proofs and Additional Theory

A.1 Unbiasedness

Let $I_i = \mathbf{1}\{i \in S\}$. Under uniform sampling of k elements without replacement, $\mathbb{E}[I_i] = k/B$. Hence

$$\mathbb{E}\left[\frac{B}{k}\sum_i I_i c_i\right] = \frac{B}{k}\sum_i \frac{k}{B} c_i = \sum_i c_i.$$

The additive bias is deterministic, which proves unbiasedness of Eq. (5). Replacing c_i by $a_{i,j} = c_{i,y^\star} - c_{i,j}$ proves the margin statement.

A.2 Fixed-cardinality variance

For scalar a_i, uniform sampling without replacement gives $\mathrm{Var}(I_i) = \frac{k}{B}(1 - \frac{k}{B})$ and, for $i \neq j$, $\mathrm{Cov}(I_i, I_j) = -\frac{k(B-k)}{B^2(B-1)}$. Writing $Z_S = (B/k)\sum_i I_i a_i$ and expanding the covariance yields

$$\begin{aligned}
\mathrm{Var}(Z_S) &= \frac{B^2}{k^2}\left[\sum_i a_i^2 \,\mathrm{Var}(I_i) + 2\sum_{i<j} a_i a_j \,\mathrm{Cov}(I_i, I_j)\right] \\
&= \frac{B(B-k)}{k(B-1)}\left(\sum_i a_i^2 - \frac{1}{B}\Big(\sum_i a_i\Big)^2\right) \\
&= \frac{B(B-k)}{k(B-1)}\sum_i (a_i - \bar{a})^2.
\end{aligned}$$

Applying this scalar identity to any logit projection proves Proposition 2. Applying it to $a_{i,j}$ proves Proposition 3.

A.3 Block interaction view

Partition h into blocks h_u and the weight vector for class c into blocks $w_{c,v}$. For sample x and class c, define $R^{x,c} \in \mathbb{R}^{B\times B}$ by $R^{x,c}_{uv} = h_u(x)^\top w_{c,v}$. Under block gauge Q, the contribution from transformed block i satisfies

$$c_i^{x,c}(Q) = [QR^{x,c}Q^\top]_{ii}.$$

Consequently $\sum_i c_i^{x,c}(Q) = \operatorname{tr} R^{x,c}$ is gauge invariant while the diagonal distribution is not. This identity provides an equivalent view of gauge selection as redistributing a fixed complete contribution across block coordinates. Only the symmetric part of R affects the diagonal of $QRQ^\top$.

B Experimental Protocol

B.1 Frozen-probe regime

We use CIFAR-100 with deterministic training, validation, and test splits and ImageNet-pretrained torchvision backbones: ResNet50, ConvNeXt-Tiny, ViT-B/16, and Swin-T. Each backbone is frozen. Three linear 100-class probes are trained independently on complete representations for each architecture and then frozen. The resulting probes have test accuracies of approximately 73.8%, 78.9%, 79.7%, and 75.3% for ResNet50, ConvNeXt-Tiny, ViT-B/16, and Swin-T, respectively. The cross-regime experiment uses $B = 8$. Separate estimator-generalization experiments also evaluate $B \in \{4, 8, 16\}$; all relevant representation dimensions are divisible by these block counts.

The original frozen-probe MaskedCE and GlobalVar gauges were not retained, so we reconstructed them deterministically under the same protocol. The probe parameter fingerprints matched exactly. Random, GlobalVar, MaskedCE, and BoundaryGauge share the same stored Q_0 for each matched run.

B.2 Native-predictor regime

We use torchvision ResNet50 (ImageNet-1K V2 weights), ConvNeXt-Tiny, ViT-B/16, and Swin-T with their original 1,000-class pretrained heads. Each model uses its weight object's official preprocessing. ImageNetV2 MatchedFrequency contains 10,000 images across the ImageNet-1K label space; a deterministic stratified split assigns 50% to calibration and 50% to held-out evaluation. Held-out top-1 accuracies are 70.04%, 71.28%, 69.70%, and 69.78%, respectively.

For BoundaryGauge and GlobalVar, no ground-truth label is accessed by the repair loss. BoundaryGauge obtains $y^\star$ and the five competitors from complete frozen logits. MaskedCE samples k uniformly from $\{2, 4, 6\}$ and four fixed-cardinality masks per batch and uses calibration labels. All methods optimize only the orthogonal gauge; backbone and task head parameters are fixed.

B.3 Gauge initialization and evaluation

Four random orthogonal Q_0 matrices per predictor are stored in float64 and reused exactly across paired methods. Learned gauges use $Q(A) = \exp(A - A^\top)Q_0$. For $B = 8$, evaluation exhaustively enumerates 28, 70, and 28 subsets at $k = 2, 4, 6$. The raw records contain one full-evaluation-set result per subset; from them we compute mean, standard deviation, 1st- and 5th-percentile, exact worst-subset, and best-subset accuracy. Native model fidelity is the fraction of partial predictions equal to the complete prediction; exact decision-flip probability is one minus fidelity.

B.4 Optimization details

All learned gauges optimize only A in $Q(A) = \exp(A - A^\top)Q_0$ while the backbone and task head remain frozen. BoundaryGauge-5 is optimized for up to 20 epochs. For BoundaryGauge-5, we consider learning rates $\{3\times10^{-3}, 10^{-2}, 3\times10^{-2}\}$ and select the learning rate and retained checkpoint using only its calibration objective. Held-out partial accuracy, 5th-percentile accuracy, worst-subset

accuracy, and the retained cardinality k are never used for model selection. BoundaryGauge and GLOBALVAR optimize objectives computed from complete representations and therefore do not sample partial-availability masks during gauge optimization. BoundaryGauge fixes $y^\star$ and its five competitors from the frozen complete logits. The supervised MASKEDCE control uses calibration labels, samples k uniformly from $\{2, 4, 6\}$, and samples four fixed-cardinality masks per batch. In the native-predictor regime, the calibration split contains 5,000 ImageNetV2 images and the disjoint held-out split contains 5,000 images. Once model selection is complete, the final gauge is frozen and evaluated by exact subset enumeration at all three availability levels.

C ADDITIONAL RESULTS

C.1 EXACT WORST-SUBSET RESULTS

For the native ResNet50, averaging the four random initializations and ordering methods as Random, GLOBALVAR, BoundaryGauge-5, and MASKEDCE, exact worst-subset accuracies are 37.50%, 35.15%, 60.43%, and 60.31% at 25% availability; 55.94%, 56.06%, 66.11%, and 66.22% at 50%; and 64.90%, 63.21%, 69.08%, and 68.94% at 75%. Thus BoundaryGauge's average improvements are accompanied by comparable gains in worst-subset robustness.

Table 1: **Detailed native ResNet50 repair.** Exact subset enumeration on ImageNetV2; learned and random rows average four matched initializations. All entries are percentages. Lower flip probability and higher worst-subset accuracy are better.

	Mean accuracy			Flip probability	Worst-subset accuracy
Gauge	25%	50%	75%	50% ↓	50% ↑
Random	46.6	61.7	67.1	24.2	55.9
GLOBALVAR	43.9	59.2	66.1	27.8	56.1
BoundaryGauge-5	**63.0**	**68.3**	**69.4**	**10.5**	66.1
MASKEDCE	62.7	68.2	69.4	10.8	**66.2**

C.2 FROZEN CROSS-ARCHITECTURE GAINS

For 25%, 50%, and 75% availability, respectively, the mean BoundaryGauge-5 gains over paired random gauges are: ResNet50, 13.10, 9.29, and 4.48 points; ConvNeXt-Tiny, 8.13, 3.22, and 1.20; ViT-B/16, 2.44, 1.71, and 0.84; and Swin-T, 3.31, 2.40, and 1.07. In all 48 combinations of architecture, probe run, and random initialization, each optimized gauge improves mean, 5th-percentile, and exact worst-subset accuracy at all three availability levels.

C.3 NATIVE LOW-HEADROOM PREDICTORS

ConvNeXt-Tiny, ViT-B/16, and Swin-T show much smaller partial-accuracy variation across gauges than the native ResNet50. Across the 36 combinations of model, random initialization, and availability level, the most negative BoundaryGauge-5 change relative to the paired random gauge is -0.071 percentage points. These predictors therefore provide a low-headroom regime in which gauge optimization has little effect.

C.4 BLOCK-COUNT GENERALIZATION OF ESTIMATOR GEOMETRY

A separate frozen-predictor experiment uses $B = 4, 8, 16$ and retained fractions 25%, 50%, and 75% for all four architectures and three independently trained probes per architecture. The variance-scaling identity in Eq. (6) is numerically verified with maximum error 6.37×10^{-12}. For the global-variance repair used in this diagnostic stage, the same final gauge improves mean accuracy at all three retained fractions in all 144 combinations of architecture, probe run, block count, and random initialization; the lower-tail criterion holds in 143 of 144. This experiment supports the separation between retained cardinality and contribution geometry; the main experiments use the

decision-aligned BoundaryGauge objective because the native ResNet50 shows that global variance alone is not sufficient for classification.

C.5 Competitor-count diagnostic

The same boundary pattern appears across several competitor sets. On the 14 native ResNet50 gauges at 50% availability, Spearman correlation with exact decision-flip probability is 0.996 for the nearest-competitor Cantelli-style risk and 1.000 for both top-5 and top-10 averaged risks. Nearest-competitor margin variance also has $\rho = 1.000$, whereas global centered-logit variance has $\rho = -0.20$. These diagnostics motivate the multi-boundary objective used in the main experiment, where $M = 5$.

D Numerical Function-Invariance Audit

All gauges analytically preserve the complete predictor in real arithmetic. In float32, different reduction and matrix-multiplication paths produce a maximum complete-logit discrepancy of 7.0572×10^{-5} on the CIFAR-100 ResNet50 setting, while complete predictions agree in 100% of cases.

We additionally reconstructed the three ResNet50 probes and matched their W, b parameter fingerprints exactly. Re-evaluating the same retained features, probe parameters, and stored gauges in float64 reduces the maximum complete-logit discrepancy to 1.6343×10^{-13}, again with 100% prediction agreement. This confirms that the larger float32 discrepancy comes from numerical accumulation rather than the compensation formula.