PAIRCAUSE: EXACT-MARGINAL DIAGNOSTICS OF MULTIMODAL READOUT AND ACCEPTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

When multimodal systems accept more errors under a common perturbation, is information lost, read out incorrectly, or simply admitted by the acceptance rule? PAIRCAUSE diagnoses these alternatives by changing the pairing of complete modality trajectories while exactly preserving their label-conditional empirical multisets. On 80 synthetic test blocks, shared pairing raises ERM accepted error from 23.32% to 46.63%, predominantly through joint-state composition; simple linear and strong-cue controls reproduce the direction. Distribution-aligned probes reveal a different phenomenon: on the same shared-mixture assessment blocks, refitting a linear head on frozen ERM features raises accuracy from 50.96% to 87.53%, but raw-input and random-feature controls perform similarly. Conversely, a clean-fitted linear head reduces Coupled's shared accuracy from 88.04% to 52.81%. Separating classification, fixed-mask risk, and recomputed acceptance shows that cue restoration combines genuine correction with rejection. Twenty retained fits and 36 new same-backend fits support these distinctions. The contribution is an auditable feature-space diagnostic of recoverability, readout supervision, and acceptance, not evidence of a uniquely learned causal representation or a demonstrated physical sensor attack.

1 INTRODUCTION

Camera–radar fusion combines complementary measurements, often through bird's-eye-view features or object queries (Lin et al., 2024; Chu et al., 2025). Agreement can help association, yet agreement alone does not certify semantic correctness. If two cues move together, their discrepancy may remain small even when their common component becomes misleading. Figure 1 separates this geometric observation from both a real-world motivation and any claim about a learned representation.

Three explanations must be distinguished. A representation may no longer contain accessible label information; a fixed output head may fail to use information that remains accessible; or an acceptance rule may admit more wrong predictions without a comparable increase in conditional risk. Probing, input restoration, and accepted-error counting address different objects. None can substitute for the others. In particular, a high probe score conditioned on an oracle perturbation state is not evidence of high unconditional decodability, and an error rejected after restoration is not necessarily an error corrected by the classifier.

We develop a controlled diagnostic rather than a new fusion architecture. Within each held-out block and label, PAIRCAUSE re-pairs entire twelve-frame trajectories, retaining every camera and radar trajectory exactly once. A paired comparison therefore changes the empirical joint input without changing either modality's evidence multiset. Fixed classifiers, thresholds, labels, and temporal endpoints make the remaining changes inspectable. Our executed study uses four-dimensional synthetic features; photographs provide scene context only.

Our contribution is the combination of an exact-marginal intervention with a support-aware descriptive decomposition, distribution-matched readout controls, and independently controlled classification and acceptance measurements. The resulting evidence is deliberately falsifiable: simple rules reproduce the headline accepted-error increase, raw inputs rival learned representations after head adaptation, and a clean-only refit can discard an augmented model's useful decision rule while

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.

leaving its hidden state frozen. These findings locate the effect more precisely than either a single robustness score or an unconditional claim of representation failure.

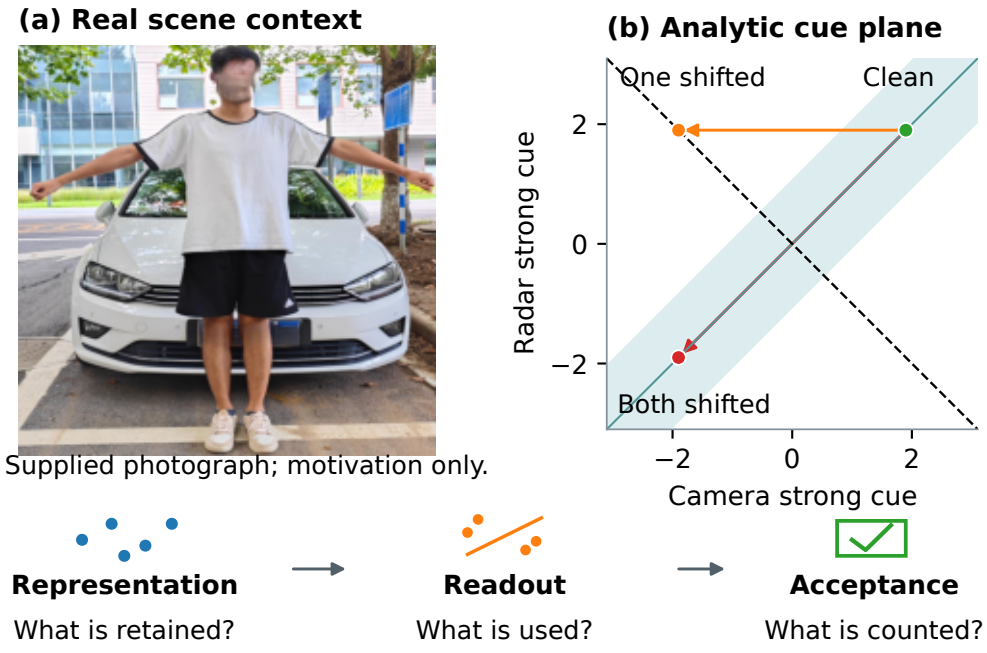


Figure 1: Agreement, representation, and acceptance are different objects. Left: a supplied historical photograph, with its existing anonymization retained, illustrates a shared scene; it is not an experimental input or synchronized trial. Right: an illustrative common shift stays inside an agreement band while crossing the rule $x_{v1} + x_{r1} = 0$; a one-channel shift leaves the band. This is an analytic schematic, not a fitted embedding or measured attack. Bottom: the three distinct diagnostic questions.

2 Exact-marginal pairing and identifiable quantities

Common and differential evidence. For aligned cues $z_v = \mu Y + a + \epsilon_v$ and $z_r = \mu Y + a + \epsilon_r$, the difference $z_v - z_r = \epsilon_v - \epsilon_r$ removes the common shift a, whereas $(z_v + z_r)/2$ retains it. This restricted cancellation argument does not imply that all classifiers fail, nor that a physical object can realize an arbitrary feature shift. An additional surviving cue can still make the class predictable.

Let V_{bci} and R_{bci} denote complete modality trajectories in block b, class stratum c, and source index i. Replacing (V_{bci}, R_{bci}) with $(V_{bci}, R_{bc\pi(i)})$ for a permutation π preserves both empirical modality multisets and every within-modality temporal sequence. In our four-trajectory stratum, each modality has two clean and two corrupted trajectories. Shared pairing S aligns their flags, giving joint counts $(2, 0, 0, 2)$ for states $00, 01, 10, 11$. Re-pairing R exchanges one clean and one corrupted radar trajectory, giving $(1, 1, 1, 1)$ (Figure 2). Labels and source blocks are unchanged; the joint inputs are not identical.

Proposition 1 (Composition under a shared indicator)**.** *If two Bernoulli corruption flags have marginal probability p and share one draw with probability λ but are otherwise independent, then*

$$\mathbb{P}(B_v = B_r = 1) = p^2 + \lambda p(1-p), \quad \mathbb{P}(B_v \neq B_r) = 2(1-\lambda)p(1-p). \tag{1}$$

At $p = 1/2$, the both-corrupted mass increases from $1/4$ to $1/2$. Our finite protocol realizes these endpoint counts exactly, rather than estimating them with fresh corruption draws. The proposition concerns joint exposure, not neural reasoning. Its proof and the distinction between a deterministic pairing diagnostic and a conditional-independence test appear in Appendix A.

Complete denominators. For classification error $E = \mathbf{1}\{\hat{Y} \neq Y\}$, confidence s, and residual D, acceptance is $A = \mathbf{1}\{s \geq \gamma\}\mathbf{1}\{D \leq \tau_D\}$. We report all-frame classification error, coverage

$C = \mathbb{P}(A = 1)$, all-frame accepted error $R = \mathbb{P}(E = 1, A = 1)$, and conditional risk $Q = R/C$ when $C > 0$. Thus $R \leq \min\{\mathbb{P}(E = 1), C\}$. A change in R does not identify a change in Q, let alone a representation mechanism. Our canonical Q divides pooled accepted-wrong counts by pooled accepted counts across fixed fits and frames; it is not an average of trajectory-wise ratios.

A decomposition without unsupported responses. Let w_{kc} be the observed mass of corruption state c in pairing k, and $e_{kc} = \mathbb{P}(EA = 1 \mid c, k)$. Re-pairing supports all four states. Shared pairing supports only 00 and 11, so its mixed-state responses remain NA. Using re-paired responses as the reference gives the exact identity

$$R_S - R_R = \underbrace{\sum_c (w_{Sc} - w_{Rc}) e_{Rc}}_{\text{reference composition}} + \underbrace{\sum_{c: w_{Sc} > 0} w_{Sc}(e_{Sc} - e_{Rc})}_{\text{supported response change}}. \tag{2}$$

No shared 01 or 10 response is imputed. The two terms depend on the reference and are descriptive standardization, not a unique causal attribution. We calculate this identity within each complete source block, then average fixed fits and blocks.

3 Executed study and diagnostic controls

Generator and sources. For $y \in \{-1, +1\}$, the four input features are

$$x_v = (2y + 0.60\epsilon_{v1},\ 0.80y + 0.90\epsilon_{v2}), \qquad x_r = (2y + 0.70\epsilon_{r1},\ 0.80y + 1.00\epsilon_{r2}). \tag{3}$$

Each noise coordinate is an independent stationary Gaussian AR(1) process with coefficient 0.65 across twelve frames. Corruption subtracts $A_b y$ from the strong coordinate, with $A_b \sim U[3, 5]$ constant within a block. The weak coordinates are unchanged. This surviving signal is an explicit assumption, not a discovered invariant. Independent source splits contain 96 training, 24 validation, 24 calibration, and 80 test blocks; a block has eight class/trajectory instances and 96 frames. The test set contains 640 trajectories and 7,680 frames, not 7,680 independent trials.

Models and thresholds. Twenty retained 4–32–32–2 ReLU networks (1,282 parameters) comprise five initializations for each of four regimes: clean ERM, modality dropout, independent strong-cue corruption, and coupled strong-cue corruption. All use 100 AdamW epochs, learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, and regime-matched validation selection. ModDrop drops one randomly selected modality with probability 0.35. Independent and Coupled have marginal corruption probability 0.35, but their both-corrupted exposure is 0.1225 and 0.35, respectively. This dose difference prevents attributing their comparison solely to learned coupling. Details and the new 36-fit same-backend repetition are in Appendix B.

The fixed confidence threshold is $\gamma = 0.8$. The feature residual is $D = (x_{v1} - x_{r1})^2/0.85$, with $\tau_D = 3.8136$ from the 95th percentile of clean calibration residuals. This is not a physical radar residual. Persistence requires at least eight accepted-wrong frames and a consecutive run of at least four within a complete twelve-frame trajectory. We distinguish *shared pairing* ($p = 0.5, \lambda = 1$) from *all-joint corruption* ($p = 1$); the latter is a different stress condition, not another name for the former.

Readout protocol. Every unconditional probe receives one frame's raw four-vector, frozen h_1, frozen h_2, or same-width randomly initialized h_2, without labels, perturbation flags, amplitude, or block identifiers as features. A single logistic readout per condition uses blocks 0–39 for fitting and blocks 40–79 for assessment, with training-only standardization and fixed $C = 1$. The original head is assessed on the same latter blocks. We also transfer clean-fitted heads to shifted conditions and fit a trajectory-shuffled-label negative control. Condition-specific readouts receive labeled examples from their fitting distribution: their improvement is supervised readout adaptation, not zero-shot robustness. Oracle state-stratified probes are kept separately and never interpreted as unconditional decoding.

Restoration protocol. We replace a corrupted strong coordinate by its own source's clean value, retaining the other coordinates, source labels, and block. The radar clean counterpart follows the

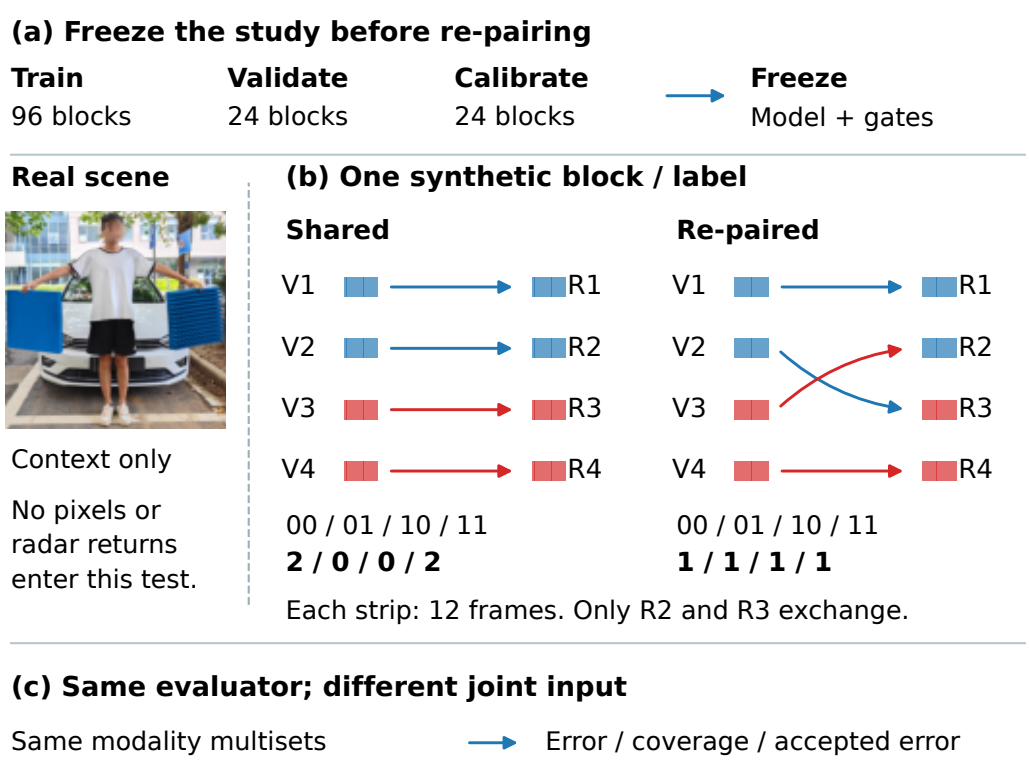


Figure 2: Change the pairing, not the modality evidence. Disjoint fitting and calibration sources are frozen first. The photograph occupies a separate context-only lane. In a synthetic block/label stratum, every twelve-frame V_i and R_i occurs exactly once in each pairing; only $R_2 \leftrightarrow R_3$ changes. Blue/red strips denote clean/corrupted strong cues, not measured sensor signals. Both pairings use the same classifier and gate thresholds.

exact re-pairing permutation. Thresholds remain fixed, but residual *values* change. For each restoration, we measure unfiltered classification error, error within the *original* accepted set, and the endpoint with all gates recomputed. An eight-cell control independently selects original/restored prediction error, confidence gate, and residual gate. The four cells that change predictions and confidence together form a 2×2 comparison against the residual source.

Uncertainty. Point estimates equally weight retained fits and equal-sized source blocks. For primary intervals we average fits within a block and resample 80 complete paired blocks 10,000 times; readout intervals use the 40 assessment blocks. Conditional-risk ratios are recomputed inside every resample. Paired changes use within-block differences, never subtraction of endpoint confidence envelopes. These intervals are conditional on the fitted models and generator. Repetition summaries separately report SD across three dataset means, each averaging three initializations.

4 What the accepted-error increase measures

Table 1 separates the endpoint's components. ERM accepted error increases by 23.31 percentage points while its coverage increases by 46.34 points; conditional risk moves only from 48.39% to 49.33%. Coupled's classification error actually decreases from 12.97% to 12.23%, even as its accepted error rises from 1.86% to 4.15%. Consequently, the accepted-error increase cannot itself demonstrate semantic information loss.

A clean-calibrated strong-sum rule and camera-only, radar-only, and four-dimensional linear controls reproduce the direction (Figure 3). In particular, the radar-only classifier's unfiltered error is invariant to radar re-pairing, yet its accepted error increases because acceptance also depends on

Table 1: Frozen-model endpoints (%). R: re-paired; S: shared pairing, both with $p = 0.5$. Each row pools five fixed fits on the same 80 blocks. C is coverage, R accepted error, and $Q = R/C$ conditional risk.

Regime	Pair	Class error	C	R	Q
ERM	R	40.61	48.20	23.32	48.39
ERM	S	49.37	94.54	46.63	49.33
ModDrop	R	43.75	48.39	23.70	48.98
ModDrop	S	49.71	94.87	47.18	49.74
Independent	R	13.44	37.49	2.58	6.87
Independent	S	14.68	72.82	5.57	7.65
Coupled	R	12.97	37.70	1.86	4.93
Coupled	S	12.23	73.45	4.15	5.65

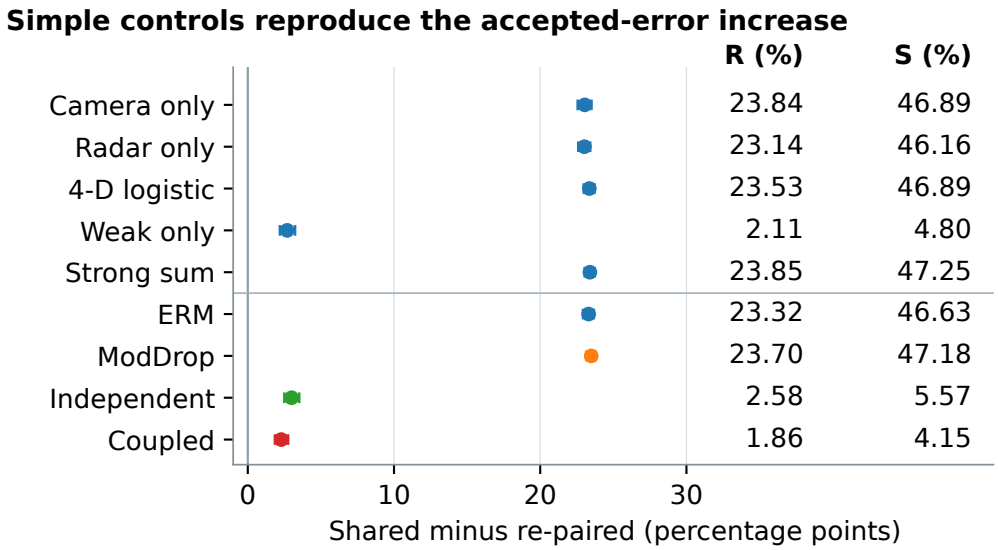


Figure 3: The direction is not neural-specific. Points and whiskers are paired accepted-error changes with 95% whole-block bootstrap intervals. The right columns give re-paired (R) and shared (S) all-frame rates. Linear baselines use clean training and validation score fitting; all models share the same fixed residual and confidence gates. Neural rows average five fixed fits within each block.

camera–radar agreement. The weak-only linear control has much lower accepted error (2.11% versus 4.80%), exposing the generator's deliberately surviving information. Baseline specifications and raw scores are retained, rather than inferred from plotted curves.

Table 2: Reference-dependent decomposition in percentage points [95% paired-block interval]. Re-paired responses define the reference; no unsupported shared response is filled in.

Regime	Composition	Response change	Total
ERM	23.09 [22.67, 23.48]	0.22 [-0.25, 0.68]	23.31 [22.93, 23.66]
ModDrop	23.36 [23.02, 23.68]	0.13 [-0.31, 0.58]	23.49 [23.15, 23.80]
Independent	2.57 [2.16, 2.98]	0.43 [-0.21, 1.08]	3.00 [2.51, 3.52]
Coupled	1.77 [1.41, 2.15]	0.52 [-0.04, 1.09]	2.29 [1.86, 2.75]

Equation 2 assigns 23.09 points of ERM's 23.31-point change to reference composition, and 0.22 points to supported response change (Table 2). The latter interval includes zero. This is stronger than an unsupported "count plus gate" explanation, but weaker than a causal mediation result. Composition changes which observed states reach a frozen classifier and evaluator; their response includes both prediction and rejection. Unsupported shared mixed states remain NA throughout. Frozen-

gate risk–coverage curves and a separately calibrated clean-90% acceptance analysis retain these denominator distinctions (Appendix G).

5 What is retained, what is read out, and what is rejected

5.1 Distribution-matched decoding identifies a readout gap, not unique encoding

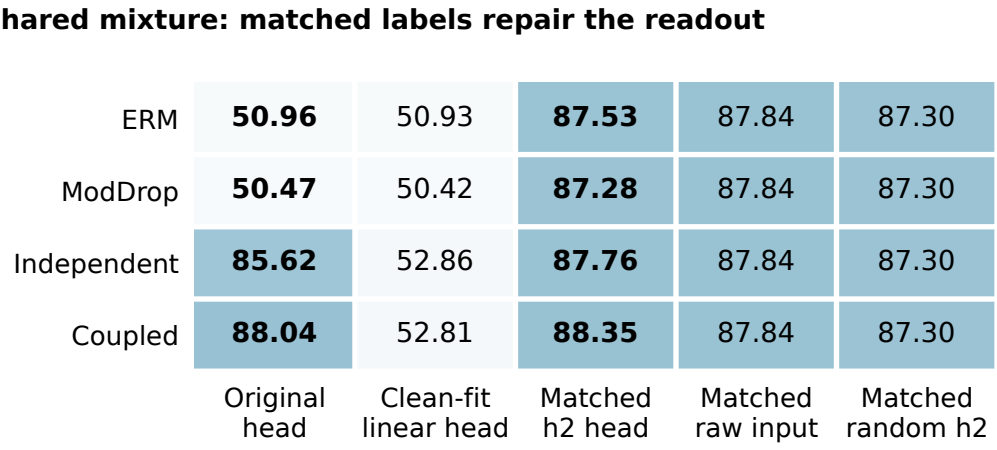


Figure 4: Readout supervision matters with the representation fixed. All columns use the same shared-mixture assessment blocks. Clean-fit and matched h_2 heads use the identical logistic specification, fitted respectively on clean or shared-mixture blocks 0–39. Raw and random-feature controls are also matched-fit. Numbers are balanced accuracies (%); no gate filters these predictions.

On the shared mixture, ERM's original head attains 50.96% balanced accuracy, versus 87.53% for a matched-fitted linear head on frozen h_2: a 36.56-point paired gain [34.89, 38.09]. ModDrop gains 36.81 points, Independent 2.14 points, and Coupled only 0.32 points [–0.34, 0.95]. Thus a large recoverable readout gap is present for the clean-trained regimes but not established for Coupled. These are common-distribution comparisons, not clean probe scores juxtaposed with corrupted model errors.

The controls narrow the interpretation. Raw-input logistic regression reaches 87.84%, and random-network h_2 reaches 87.30% on the same shared assessment blocks. Neither requires a learned fusion representation. The trajectory-shuffled-label controls are near chance (48.67–49.39% across regimes). We therefore infer accessible label signal and a readout mismatch, not a specially learned causal code or an information-theoretic advantage of hidden states. Probe failure would also not, by itself, prove information absence.

Conversely, a *clean-fitted* linear head obtains only 50.93% on ERM and 52.81% on Coupled shared features. Coupled's native head achieves 88.04% without that refit. Changing head supervision can therefore remove useful decision behavior while leaving the learned representation fixed. Cross-condition transfer matrices in Appendix D make clear that a successful matched refit is not automatically transferable.

Unconditional h_2 label decoding is 87.28–88.35% on the shared mixture, not nearly 100%. Near-perfect oracle scores fit separate classifiers *within known corruption states*; their conditioning supplies information unavailable to the deployed head. Our unconditional perturbation-state probes attain approximately 83–88%, rather than over 98%. For an independent analytic reference, the generator's equal-prior, single-frame shared-mixture Bayes accuracy is approximately 89.35% when state and block amplitude are unknown. This population benchmark is condition-specific and is not a hard upper bound on every finite test realization (Appendix C).

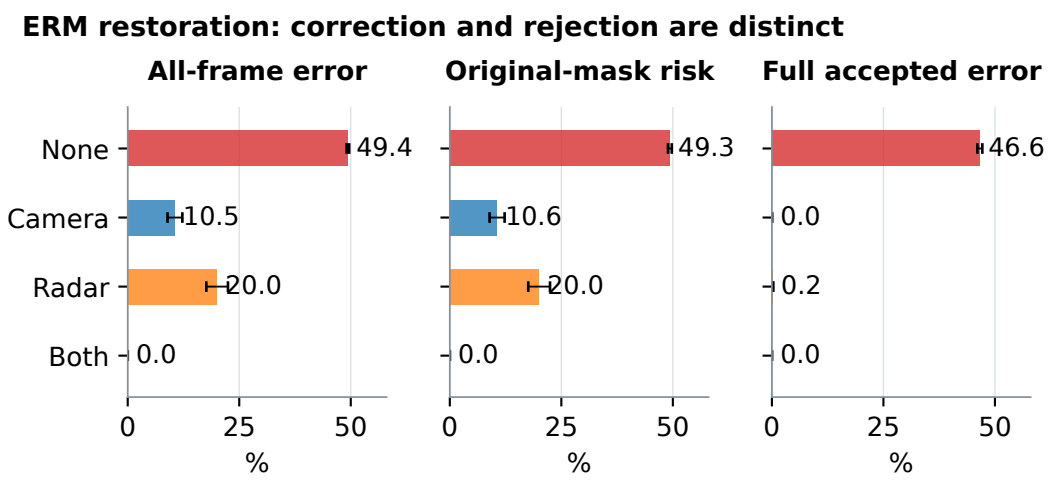


Figure 5: Three distinct consequences of source-matched restoration. ERM under shared pairing: left, classification error on all frames; middle, conditional error on the original accepted set, held fixed; right, all-frame accepted error after recomputing confidence and residual gates. Bars restore no cue, camera, radar, or both. Whiskers are 95% paired-block intervals. Near-zero accepted error is not near-zero classification error.

5.2 Restoration combines classification correction with rejection

Camera restoration on shared inputs reduces ERM's unfiltered error from 49.37% to 10.54%, and its error within the original accepted set from 49.33% to 10.56%. These changes establish classification correction for this source-defined intervention. Yet the recomputed accepted-error rate falls further, to 0.02%, while coverage falls from 94.54% to 48.32%. Rejection therefore accounts for an important part of the final endpoint's apparent success. Radar restoration leaves 20.02% unfiltered error despite only 0.21% accepted error. Restoring both yields zero observed ERM classification errors on these sources, not a population-risk guarantee.

The same separation holds for re-paired inputs: camera restoration leaves 10.29% classification error and 9.59% error on the original accepted set even when accepted error is zero. The eight-cell diagnostic reports all prediction, confidence, and residual combinations (Appendix E). As a decisive negative control, camera restoration leaves every radar-only prediction exactly unchanged, while shared accepted error falls from 46.16% to 0.44%. That improvement is rejection, not a corrected classifier.

5.3 Margin interaction is gauge-invariant but scale-dependent

For binary logits z_1, z_0, define $m = z_1 - z_0 = \log(p_1/p_0)$ and

$$J_m = m(v_1, r_1) + m(v_2, r_2) - m(v_1, r_2) - m(v_2, r_1). \tag{4}$$

Adding any common input-dependent function to both logits preserves m, probabilities, and J_m. A single-class-logit analogue does not have this invariance (Appendix A). Each four-way tuple uses the same block, class, and frame and every clean/corrupted source pair. We additionally divide $|J_m|$ by the clean-calibration SD of each fitted model's margin before averaging.

Raw mean $|J_m|$ is 3.508, 5.799, 0.415, and 1.139 for ERM, ModDrop, Independent, and Coupled. The normalized values are 0.202, 0.274, 0.091, and 0.263; raw magnitude alone would obscure the proximity of ModDrop and Coupled on that scale. An additive strong-cue rule has zero interaction to numerical precision. Nonzero J_m diagnoses nonadditivity, not causal reasoning or information loss. In particular, substantial normalized interaction coexists with Coupled's negligible matched-readout gap.

6 REPETITION, RELATED WORK, AND SCOPE

Table 3: Same-backend repetition: mean ± SD over three dataset means, each averaging three starts. Both quantities are percentage-point changes; the head gain uses shared-mixture assessment blocks 40–79.

Regime	Accepted-error increase	Matched h_2 head gain
ERM	23.42 ± 0.04	37.08 ± 0.10
ModDrop	23.48 ± 0.03	37.18 ± 0.16
Independent	3.09 ± 0.11	2.25 ± 0.21
Coupled	2.32 ± 0.05	0.12 ± 0.13

Beyond one training realization. We newly train three independent training/validation realizations with three initializations per regime using the same PyTorch AdamW and checkpoint-selection procedure as the primary study: 36 fits, distinct from the 20 retained fits. Every fit receives the endpoint, margin, restoration, and unconditional readout checks. Table 3 reports between-dataset variation without treating nine starts as nine independent datasets. ERM and ModDrop retain large readout gaps; Coupled does not. Camera-restoration classification error and normalized interaction vary more than the headline endpoint, as detailed in Appendix F. The common test source and synthetic generator still limit generalization. The older 36-fit NumPy auxiliary control is preserved separately because its actual update rule differs, not merely its floating-point backend.

Relation to existing diagnostics. Conditional permutation tests use a specified conditional resampling law and assumptions to control null-test error (Berrett et al., 2019); exact preservation of empirical trajectory multisets alone supplies no such guarantee. Our object is a paired behavioral contrast. Selective prediction already distinguishes coverage from risk (Geifman & El-Yaniv, 2019); we do not claim that $R = CQ$ is new, but use those denominators to avoid misattributing an acceptance effect to learned representations. Probe-control work distinguishes information accessibility from the probe's own learning (Hewitt & Liang, 2019); our raw-input, random-feature, clean-fit-transfer, and shuffled-label controls operationalize that concern for this generator, rather than reproducing their linguistic control task.

Fusion architectures and uncertainty-aware robustness methods motivate the application (Lin et al., 2024; Chu et al., 2025; Cho et al., 2025), while shortcut learning provides a broader account of distribution-dependent decision rules (Geirhos et al., 2020). These are related methods, not evaluated real-sensor baselines. A four-dimensional feature model cannot substitute for their pixel, point-cloud, or object-query representations.

Limits and conclusion. This study isolates input coupling, accessible signal, readout supervision, and evaluator selection in a small, explicitly specified system. It does not establish a neural-only failure, uniquely causal features, a physical intervention, or universal fusion behavior. Weak-cue corruption stress tests show dependence on the surviving-cue assumption; Independent/Coupled retain their exposure confound. No synchronized real-sensor recording and runnable frozen fusion encoder were supplied or evaluated. The photographs are excluded from all numerical denominators. Within this boundary, exact-marginal pairing makes a useful distinction possible: the system can accept more errors even when recoverable signal remains, and readout adaptation can improve predictions without altering the frozen features. Reporting these objects together prevents a coverage change from being mistaken for information loss.

ETHICS STATEMENT

The executed experiments use synthetic four-dimensional feature trajectories and do not interact with vehicles, sensor hardware, public traffic, or new human participants. The paper does not demonstrate a commercial-product vulnerability or a physically realizable attack. The historical photographs were supplied with the source archive; their existing face and license-plate obfuscation is retained. They are used for motivation only and excluded from model inputs and numerical evidence. The archive does not supply publication-rights or consent documentation; the authors must confirm those permissions before public submission. Physical replication would require authorization, controlled sites, and protection of bystanders.

REPRODUCIBILITY STATEMENT

The artifact contains twenty retained primary checkpoints, thirty-six newly trained same-backend checkpoints, disjoint source data, saved unconditional probe heads and predictions, all three-way restoration-control combinations, margin and additive-null checks, block-level sufficient statistics, and figure-generation code. Appendix H specifies entry points and evidence paths. Older NumPy auxiliary fits and physical aggregates remain separately labeled historical records, not replacements for the current executions. The numerical PDF figures are generated by the included Python renderer. The accompanying single MATLAB file contains the exported numerical ordinates in paper order; native MATLAB execution was unavailable and is not claimed. No raw synchronized sensor recording or runnable real fusion encoder was supplied or evaluated.

AI USE STATEMENT

An OpenAI assistant substantially assisted problem formulation, manuscript restructuring and drafting, source-code inspection, mathematical checks, implementation, synthetic training and evaluation, statistical analysis, plotting, tests, and document checks. Reported new numerical results come from retained executions; no physical measurements, real-sensor model performance, image permissions, or human approvals were generated. Figures 1 and 2 combine directly extracted supplied photographic assets with newly drawn editable vector diagrams; the scene pixels are not AI-generated. The authors are responsible for independently verifying the scientific claims, source provenance, code, citations, permissions, and final submission.

REFERENCES

Thomas B. Berrett, Yi Wang, Rina Foygel Barber, and Richard J. Samworth. The conditional permutation test for independence while controlling for confounders. arXiv:1807.05405, version 2, 2019. URL https://arxiv.org/abs/1807.05405.

Minkyoung Cho, Yulong Cao, Jiachen Sun, Qingzhao Zhang, Marco Pavone, Jeong Joon Park, Heng Yang, and Zhuoqing Mao. Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6580845ce91a748a9bb09cec079aac1d-Abstract-Conference.html.

Xiaomeng Chu, Jiajun Deng, Guoliang You, Yifan Duan, Houqiang Li, and Yanyong Zhang. RaCFormer: Towards high-quality 3d object detection via query-based radar-camera fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '25, pp. 17081–17091. IEEE/CVF, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chu_RaCFormer_Towards_High-Quality_3D_Object_Detection_via_Query-based_Radar-Camera_Fusion_CVPR_2025_paper.html.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. RCBEVDet: Radar-camera fusion in bird's eye view for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '24, pp. 14928–14937. IEEE/CVF, 2024. doi: 10.1109/CVPR52733.2024.01414. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_RCBEVDet_Radar-camera_Fusion_in_Birds_Eye_View_for_3D_Object_CVPR_2024_paper.html.

A IDENTITIES, INVARIANCES, AND INFERENTIAL LIMITS

Proof of the shared-indicator identity. In the independent branch, the both-corrupted probability is p^2 and the disagreement probability is $2p(1-p)$. In the shared branch they are p and zero. Their mixture gives $(1-\lambda)p^2 + \lambda p = p^2 + \lambda p(1-p)$ and $2(1-\lambda)p(1-p)$. Each marginal remains Bernoulli(p). In the executed endpoint pair, the exact finite counts are imposed by a within-block, within-label permutation, not by asymptotically approximating independent flag frequencies.

What exact marginal preservation guarantees. A permutation of complete radar trajectories is a bijection, so every function of the radar multiset alone is unchanged, as is every function of the unchanged camera multiset alone. This includes label-conditional single-modality classification error for a frozen unimodal model. It does not include a cross-modal acceptance rule. The equality applies to complete trajectories, not only their means: individual frame re-pairing would preserve frame marginals but could destroy temporal sequences and change persistence. We check full-array trajectory identities. Selecting a particular permutation with the desired joint-state counts is not sampling a valid null law for a conditional-independence test; the latter requires additional distributional assumptions (Berrett et al., 2019).

Support-aware standardization. Write $R_k = \sum_c w_{kc} e_{kc}$. Add and subtract $\sum_c w_{Sc} e_{Rc}$ to obtain Equation 2. Every e_{Rc} is observed because re-pairing has all four cells. Terms with $w_{Sc} = 0$ vanish, so the second sum needs only shared 00 and 11. The reverse reference would require unobserved shared responses for re-paired mixed cells and is not used. The weights are $w_R = (1/4, 1/4, 1/4, 1/4)$ and $w_S = (1/2, 0, 0, 1/2)$ in every block and class. Per-block totals and both components reconcile to numerical precision; confidence intervals need not add endpoint-wise because the components are correlated.

Why a single logit is not a decision interaction. For any $g(x)$, replace z_1, z_0 by $z_1 + g, z_0 + g$. Softmax probabilities, predicted labels, and confidence are unchanged, but a single-logit interaction changes by $J[g]$. Setting $g = -z_1$ makes the class-1 logit and its interaction identically zero. The transformation can be implemented by subtracting the original class-1 affine output from both outputs; no extra hidden capacity is needed. In contrast, the margin and Equation 4 are unchanged. Our checks subtract the evaluated class-1 logit from both logits in float64; margin discrepancies are zero to numerical precision for all 56 current fits. The additive rule $m = x_{v1} + x_{r1}$ gives $\max |J_m| < 10^{-12}$ on the same tuples.

Scale normalization is limited. We compute $\sigma_m = \mathrm{SD}_{\text{clean calibration}}(m)$ separately per fit and report $\mathbb{E}|J_m|/\sigma_m$. Under positive global rescaling of margins, both numerator and denominator scale together. This does not provide invariance under arbitrary recalibration, does not turn interaction into an information measure, and does not rank models by causal understanding. The calibration set is never chosen using test interaction values.

B SOURCES, FITTING, AND EXACT EXECUTION INVENTORY

The generator draws separate stationary Gaussian AR(1) processes for every block, label, trajectory, modality, and feature. It starts with a standard normal and updates $\epsilon_t = 0.65\epsilon_{t-1} + \sqrt{1 - 0.65^2}\eta_t$. Each block has two label strata, four trajectories per stratum, and twelve frames. The strong and weak features follow Equation 3. Within every test block/label, the corruption flags are a seeded permutation of $(0, 0, 1, 1)$. Re-pairing exchanges the second clean radar index with the first corrupted radar index and applies that same permutation to the clean radar counterpart used for restoration.

All networks are affine 4–32–32–2 with two ReLU layers and binary softmax. No normalization is fitted for network inference. Inputs contain only $[x_{v1}, x_{v2}, x_{r1}, x_{r2}]$, never a class, flag, amplitude, or source identifier. The primary initialization seeds are 11, 22, 33, 44, and 55. Repeat initialization seeds are 101, 202, and 303, reused within each of three data realizations. AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, and 100 epochs (3,600 minibatch updates). All epochs execute; the minimum regime-matched validation cross-entropy selects the checkpoint. A validation augmentation fixture uses generator seed 88000 and is frozen across epochs. Its distribution differs between regimes, so validation loss is not a common robustness ranking.

Table 4: Source splits and seeds. Conditions reuse the same held-out test sources; they do not add independent scenes.

Source	Generator seed	Blocks	Frames
Primary training	12001	96	9,216
Primary validation	22001	24	2,304
Fixed clean calibration	32001	24	2,304
Fixed assessment source	42001	80	7,680
Repeat training (each)	52001–52003	96	9,216
Repeat validation (each)	152001–152003	24	2,304

Augmentation amplitudes are independently sampled from $U[3, 5]$ per training example. Independent samples two Bernoulli(0.35) flags and Coupled shares one; both subtract a class-signed amplitude from the corresponding strong feature. ModDrop sets a uniformly selected whole modality to zero with probability 0.35. Training shuffles frames for optimization, but statistical inference never treats frames as independent experimental units. The repeated fitting script imports the same augmentation implementation as the primary training script.

The new repeat execution used PyTorch 2.10.0+cpu, NumPy 2.3.5, one CPU thread, and deterministic algorithms. Selected weights are exported as NumPy arrays; validation probabilities agree with native Torch to within 10^{-6} for all 36 new fits. The current analysis reloads all 56 checkpoints. Primary probabilities agree with the retained predictions to a maximum absolute difference of 1.20×10^{-7}. Original checkpoints and original primary training data are not overwritten.

The historical 36-fit NumPy auxiliary experiment remains in its original directory. Source inspection finds an epoch-indexed Adam bias correction inside the minibatch loop and a weight-gradient L2 term in addition to decoupled decay. These are algorithmic differences, not just floating-point variation. Its existing endpoint summary reports SD over nine fits per regime. It is not pooled into the current same-backend repetitions, which instead report SD over three dataset means. This preserves the older evidence without claiming equivalence it does not have.

C A DISTRIBUTION-SPECIFIC BAYES REFERENCE

Consider a single frame with equal class prior, shared mixture $p = 1/2, \lambda = 1$, unknown corruption state, and unknown block amplitude $A \sim U[3, 5]$. The model receives exactly four features and no neighboring frames or other trajectories. The strong-coordinate sufficient statistic is the precision-weighted average

$$S = \frac{x_{v1}/0.60^2 + x_{r1}/0.70^2}{1/0.60^2 + 1/0.70^2}, \qquad \sigma^2 = \left(1/0.60^2 + 1/0.70^2\right)^{-1}. \tag{5}$$

The orthogonal Gaussian residual contains no class information conditional on this model's shared strong mean. For $Y = +1$, the clean mean is 2 and the corrupted mean $2-A$ is uniform on $[-3, -1]$. Consequently, with ϕ_σ a centered Gaussian density and Φ the standard normal CDF,

$$f_+(s) = \tfrac{1}{2}\phi_\sigma(s-2) + \tfrac{1}{4}\left[\Phi((s+3)/\sigma) - \Phi((s+1)/\sigma)\right], \quad f_-(s) = f_+(-s). \tag{6}$$

The weak-cue log likelihood ratio is $L_w = 2(0.8)(x_{v2}/0.9^2 + x_{r2}/1^2)$. Conditional on $Y = +1$, it is Gaussian with mean $2k$ and variance $4k$, where $k = 0.8^2(1/0.9^2 + 1) = 1.4301234568$. It is independent of the strong cues. Symmetry therefore gives the population Bayes accuracy

$$\mathcal{A}^* = \int f_+(s)\, \Phi\left(\frac{2k + \log f_+(s) - \log f_-(s)}{2\sqrt{k}}\right) ds. \tag{7}$$

Numerical quadrature yields 0.8935034084 (estimated numerical integration error 2.9×10^{-9} on $[-12, 12]$; negligible Gaussian tails). This is not a claim that any finite held-out accuracy above 89.35% is impossible. It does not apply to clean data, all-joint data with known condition, oracle state-stratified probes, trajectory aggregation, or an observer allowed to infer amplitude from multiple same-block examples. Its role is to align the information available to the predictor before interpreting near-perfect probe scores.

D UNCONDITIONAL READOUTS AND DISTRIBUTION TRANSFER

Every readout is logistic regression with intercept, $C = 1$, class balancing, the L-BFGS solver, tolerance 10^{-7}, and at most 1,500 iterations. Standardization is fitted only on the 3,840 frames in blocks 0–39; assessment uses the separate 3,840 frames in blocks 40–79. Each block is exactly label-balanced, so block accuracy equals balanced accuracy. These hyperparameters are fixed, not selected on assessment labels. No trajectory aggregation occurs. The raw-input readout has four input coordinates; the hidden and random-feature readouts have 32. Random features use the same architecture with initialization seeds matched to the fitted models and no training.

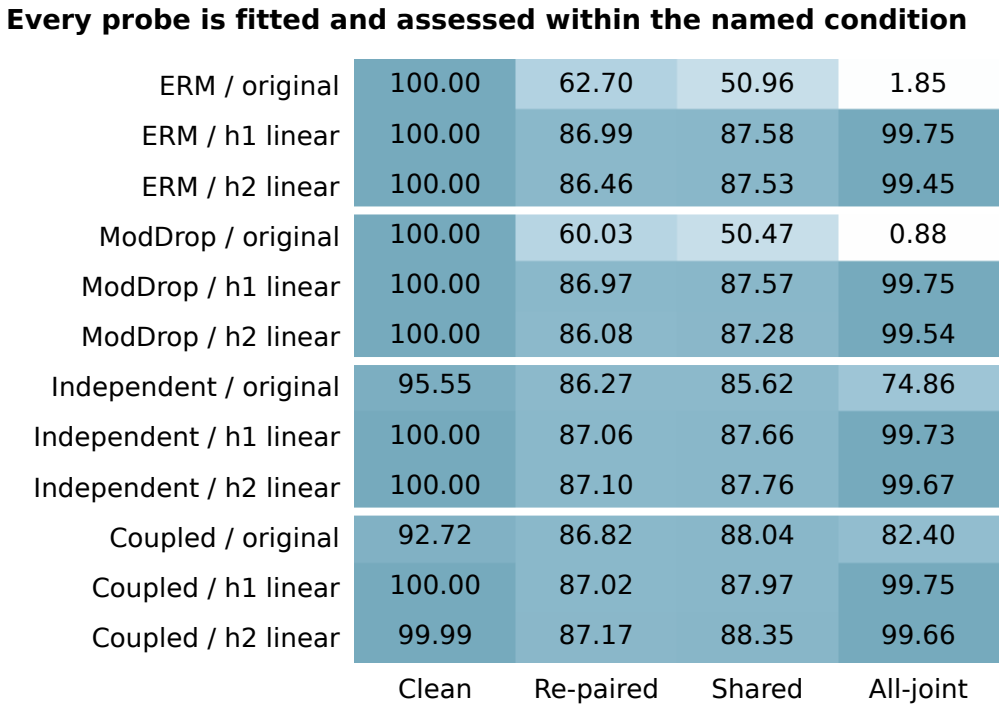


Figure 6: Decoding must be indexed by distribution. Original-head and matched linear-readout balanced accuracy (%) on the same 40 assessment blocks. Clean and all-joint $p = 1$ conditions can be decoded nearly perfectly after condition-matched fitting; neither result implies near-perfect decoding of the unknown-state shared mixture. Re-paired/shared both have $p = 0.5$.

The negative control shuffles four positive and four negative trajectory labels within each block, then repeats each assigned label over the trajectory's twelve frames. Independent block assignments prevent train/assessment source overlap while retaining the temporal label structure. Near-chance scores on this control are useful checks, not a proof of causal semantics or an exact reproduction of word-type control tasks in Hewitt & Liang (2019). Perturbation-state probes also use one unconditional readout; they do not condition on true class or the other flag. In the shared condition the two modality flags are identical, so their decoding accuracies are necessarily identical.

The retained older class probes trained one classifier per known joint corruption state. Their individual flag probes additionally conditioned on class and the other flag. Those are legitimate conditional diagnostics, but they answer an oracle-assisted question. Their original outputs and methods remain available in the historical mechanism directory. The current manuscript does not use their near-perfect scores as evidence for unconditional label retention or unconditional state recovery.

Clean-trained readouts transfer poorly to the shared mixture in every regime, including those whose original heads are robust there. Matched readouts exploit available labels from the shifted fitting condition and can transfer between some, but not all, corruption distributions. The comparison

Table 5: Shared-mixture balanced accuracy and paired head gain [95% interval]. Accuracy is in percent; gain is in percentage points. All use the same 40 assessment blocks.

Regime	Original head	Matched h_2 head	Paired gain
ERM	50.96 [50.51, 51.48]	87.53 [86.02, 88.90]	36.56 [34.89, 38.09]
ModDrop	50.47 [50.20, 50.79]	87.28 [85.73, 88.69]	36.81 [35.14, 38.33]
Independent	85.62 [84.16, 86.98]	87.76 [86.19, 89.22]	2.14 [1.00, 3.23]
Coupled	88.04 [86.73, 89.27]	88.35 [86.74, 89.83]	0.32 [-0.34, 0.95]

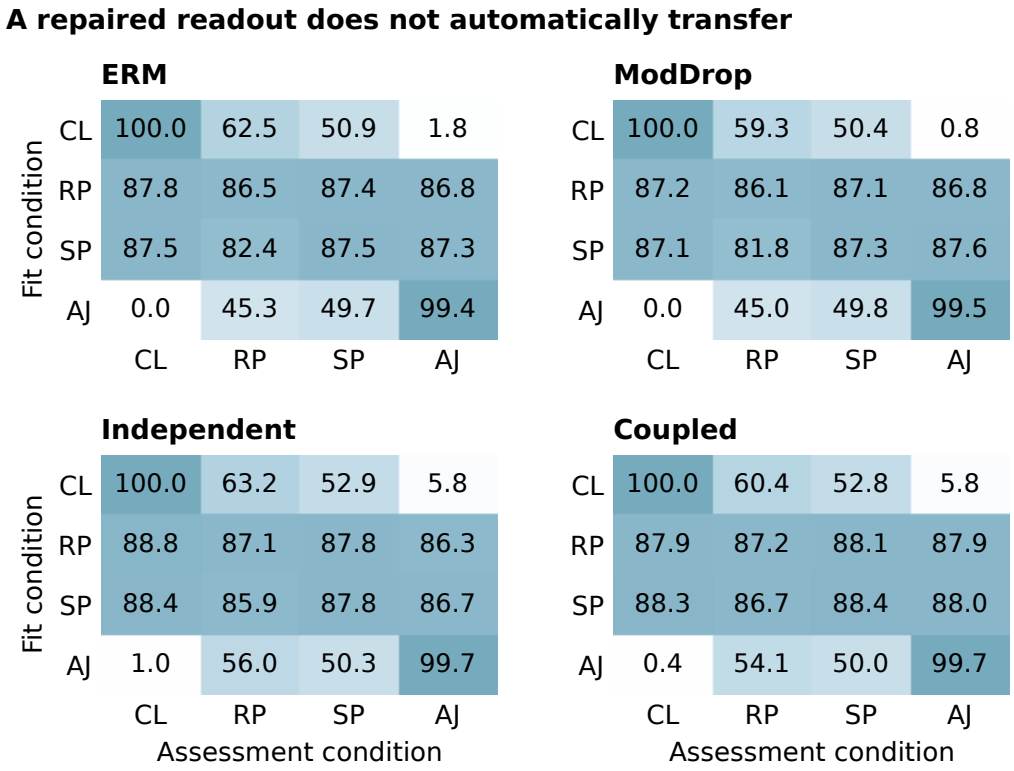


Figure 7: Frozen-feature head transfer. Rows give the fitting condition and columns the assessment condition of the same saved h_2 linear head. CL: clean; RP: re-paired $p = 0.5$; SP: shared $p = 0.5$; AJ: all-joint $p = 1$. Values are balanced accuracies (%). Each head's fitting and assessment source blocks are disjoint even when conditions differ. Diagonal entries match the unconditional readout analysis.

isolates readout training while holding the representation fixed; it does not assert that the matched head would be available at deployment without labeled adaptation data.

E Restoration and independent gate controls

Let E^0, C^0, D^0 be original error, confidence-pass, and residual-pass indicators, and let superscript 1 denote recomputation after the source-matched restoration. We save all eight rates $\mathbb{E}[E^i C^j D^k]$, $i, j, k \in \{0, 1\}$. The four cells with $i = j$ compare original/restored prediction–confidence packages against original/restored residual gates. The full eight cells additionally separate changed class decisions from changed confidence. These are controlled evaluations of stored deterministic functions, not estimated causal mediation effects.

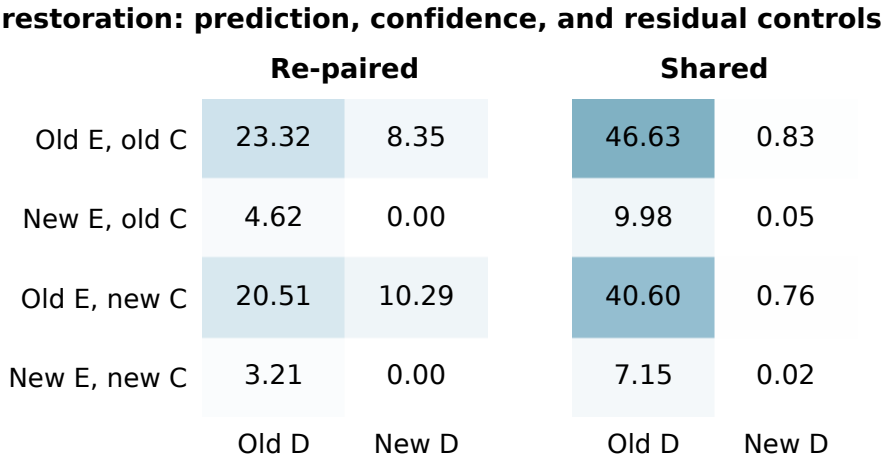


Figure 8: Camera restoration with independent indicator sources. Values are all-frame accepted-error percentages for ERM. E is the prediction error indicator, C the confidence gate, and D the residual gate. Old/new select the pre-/post-restoration indicator. The first cell reproduces the original endpoint; the last reproduces the fully recomputed endpoint. Mixing indicator sources is a diagnostic control, not a deployable sensing rule.

The fixed-mask conditional metric is $\sum E^1 C^0 D^0 / \sum C^0 D^0$. Its denominator is explicitly the original accepted set. The full-system conditional risk instead divides $\sum E^1 C^1 D^1$ by $\sum C^1 D^1$. Unfiltered error divides $\sum E^1$ by all frames. These three denominators must not be conflated. Every bootstrap resample recomputes the relevant ratio, including its fixed or recomputed denominator.

Table 6: A classification-invariant negative control (%). Camera restoration leaves every radar-only score unchanged, but changes the residual acceptance rule.

Pairing	Camera intervention	Class error	Coverage	Accepted error
RP	Original	48.84	48.03	23.14
RP	Camera restored	48.84	48.01	0.42
SP	Original	48.84	94.21	46.16
SP	Camera restored	48.84	48.20	0.44

The radar-only control is fitted on clean training data with a validation-fitted score map. Restoring camera changes no radar feature, and every radar-only score is exactly identical before and after restoration. Its large accepted-error reduction therefore comes entirely from the residual gate. This is a concrete counterexample to inferring classifier correction solely from a successful restoration endpoint. All neural restoration results for both pairings and all four regimes, including signed true-class margin changes and prediction flips, are in the current CSV and block-array outputs.

F Interaction scale and repeated mechanism evidence

For each block and class, the interaction analysis enumerates all 2×2 clean/corrupted source-trajectory choices and compares the four modality combinations at each frame. There are eight source pairs per block, 640 source pairs total, and 7,680 frame-level interaction evaluations per fitted model. Each evaluation is assigned to its original block for uncertainty; the four terms are not treated as independent samples. The retained raw logits already encode the binary margin calculation. Recomputed model-level interaction means differ from the retained margin ordinates by less than 2×10^{-15} when those stored logits are accumulated in float64.

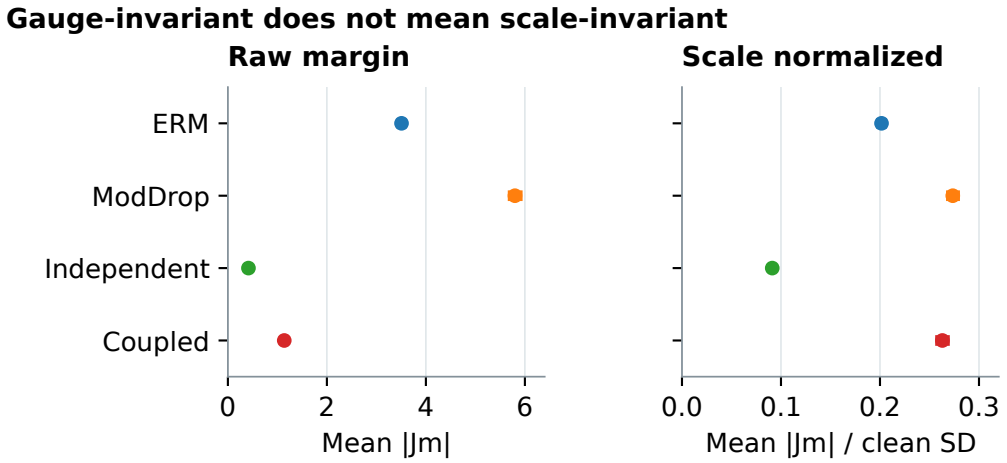


Figure 9: Interaction depends on scale as well as nonadditivity. Raw mean absolute margin interaction and the same quantity divided by each fit's clean-calibration margin SD. Whiskers resample the 80 source blocks with five fits fixed. Both definitions are invariant to a common additive logit offset; only the normalized definition removes a positive global margin rescaling. Neither is a causal-reasoning score.

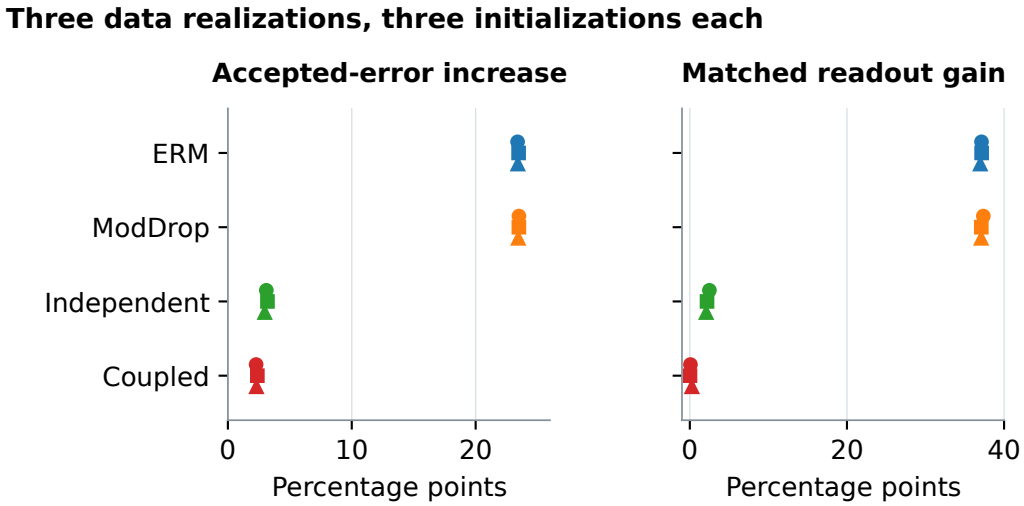


Figure 10: Independent source-realization repetition. Each circle, square, or triangle averages three initializations for one of the three newly sampled training/validation datasets. Left: shared-minus-re-paired accepted error; right: matched h_2 readout minus original-head accuracy on shared assessment blocks. All points are in percentage points and use the same fixed test source. Symbols do not imply nine independent datasets.

The endpoint direction and large clean-trained-model readout gap are stable across these three realizations. Mechanistic magnitudes are less stable: for example, repeated ERM camera-restoration classification error has SD 2.03 percentage points across dataset means, whereas its accepted-error increase has SD 0.04 points. Repeated raw and normalized interactions also vary. We therefore report repetition of each diagnostic rather than inferring mechanism stability from endpoint stability. Three training/validation realizations remain a small sample, and all reuse the same synthetic generator and assessment source.

Table 7: Repeated mechanism diagnostics: mean ± SD across three dataset means. Restoration columns refer to camera restoration on shared $p = 0.5$ inputs; both errors are percentages.

Regime	Raw $\lvert J_m \rvert$	Normalized	Class error	Accepted error
ERM	3.38 ± 0.55	0.19 ± 0.03	9.99 ± 2.03	0.02 ± 0.01
ModDrop	4.36 ± 1.08	0.21 ± 0.01	16.89 ± 3.99	0.07 ± 0.02
Independent	0.39 ± 0.01	0.08 ± 0.00	8.07 ± 0.35	0.70 ± 0.06
Coupled	1.21 ± 0.03	0.28 ± 0.02	8.32 ± 0.54	1.33 ± 0.12

G ACCEPTANCE, ESTIMANDS, AND ASSUMPTION STRESS TESTS

Table 8: Complete confidence, residual, coverage and persistence rates (%). CL: clean; RP/SP: $p = 0.5$ re-paired/shared; AJ: $p = 1$ all-joint. Persistence counts complete trajectories, not frames.

Regime	Input	Conf. pass	Resid. pass	Coverage	Persistence
ERM	CL	100.00	95.08	95.08	0.00
ERM	RP	91.96	48.67	48.20	24.47
ERM	SP	99.43	95.08	94.54	49.16
ERM	AJ	98.85	95.08	94.01	98.19
ModDrop	CL	100.00	95.08	95.08	0.00
ModDrop	RP	92.98	48.67	48.39	24.84
ModDrop	SP	99.79	95.08	94.87	49.69
ModDrop	AJ	99.58	95.08	94.68	99.06
Independent	CL	88.11	95.08	83.97	0.00
Independent	RP	76.99	48.67	37.49	0.00
Independent	SP	76.39	95.08	72.82	0.00
Independent	AJ	66.17	95.08	63.03	0.50
Coupled	CL	83.50	95.08	79.47	0.00
Coupled	RP	77.65	48.67	37.70	0.00
Coupled	SP	77.14	95.08	73.45	0.00
Coupled	AJ	71.61	95.08	68.18	0.00

The residual pass rate is model-independent and is not plotted as four separate overlapping curves. Its re-paired/shared values are 48.67% and 95.08%. Confidence pass is model-dependent; full coverage is the intersection of both gates and cannot be replaced by the residual pass rate. A zero observed persistent-event cell produces a degenerate empirical bootstrap interval, not a distribution-free zero-risk certificate. The block counts, all-frame errors, and exact temporal rule remain available.

Table 9: Different estimands from the same retained predictions (%). The legacy column averages valid trajectory ratios within each block, then blocks and fits. The canonical column divides pooled counts.

Regime	Input	Legacy traj. mean	Block-ratio mean	Pooled ratio
ERM	RP	46.97	48.47	48.39
ERM	SP	49.56	49.32	49.33
ModDrop	RP	47.74	49.06	48.98
ModDrop	SP	49.81	49.75	49.74
Independent	RP	8.37	6.95	6.87
Independent	SP	9.82	7.69	7.65
Coupled	RP	6.32	5.02	4.93
Coupled	SP	7.14	5.69	5.65

The historical conditional-risk change is explained by an estimand change, not by changed classifications. In particular, ERM re-paired 46.97% is the equal-block average of within-block means of valid *trajectory* conditional ratios. The pooled count ratio is $8{,}956/18{,}508 = 48.39\%$. Even the mean of *block* conditional ratios is different (48.47%). Current point estimates, error bars, and

resamples consistently use pooled counts. With equal block sizes, unconditional frame-average metrics are unaffected by these denominator choices.

Table 10: Clean-90%-matched acceptance control. Test clean coverage and accepted error (%) use separate denominators. SP is $p = 0.5$; AJ is $p = 1$. Thresholds are chosen only on calibration, implemented in margin space to avoid probability saturation.

Regime	Clean coverage	SP accepted error [95%]	AJ accepted error [95%]
ERM	89.65	29.24 [26.00, 32.50]	59.09 [52.72, 65.46]
ModDrop	89.31	33.74 [30.97, 36.48]	67.58 [62.22, 72.92]
Independent	89.67	8.93 [8.12, 9.74]	15.88 [14.58, 17.19]
Coupled	89.91	8.35 [7.58, 9.12]	12.24 [11.13, 13.37]

The matched-acceptance control first selects frames passing the fixed residual gate on clean calibration. A model-specific absolute-margin threshold κ is chosen so at least $\lceil 0.90 N_{\text{cal}} \rceil$ total calibration frames pass both conditions. Inclusive ties are retained. Acceptance at test is $|m| \geq \kappa$ and $D \leq \tau_D$, equivalent to a confidence threshold $\text{sigmoid}(\kappa)$. Margin-space implementation avoids rounding extremely confident float32 probabilities to one. The slight change from older matched-score ordinates is documented, not silently identified as the same numerical rule. The old ERM 59.19% "shared wrong" column refers to $p = 1$ all-joint corruption; the current margin-space recalculation is 59.09%. Neither is the $p = 0.5$ shared-pairing condition whose classification error is 49.37%.

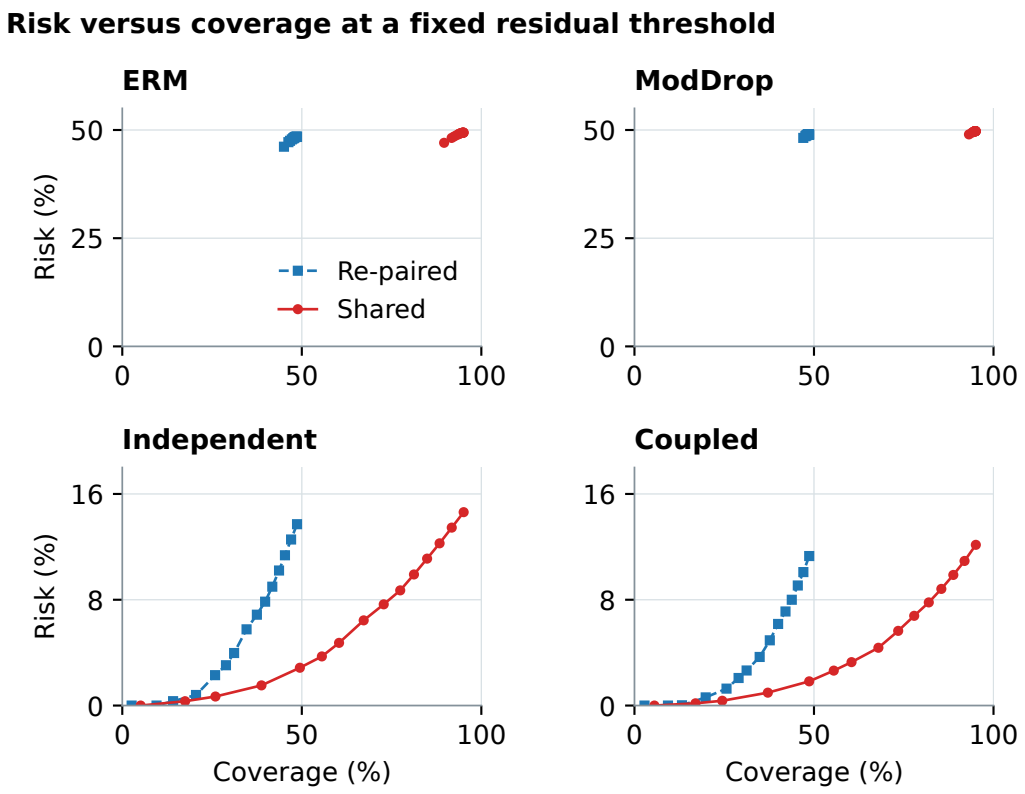


Figure 11: Risk–coverage curves with frozen residual calibration. The score-threshold grid is 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.925, 0.95, 0.975, 0.99, 0.995, and 0.999; it is not optimized on the test set. Curves show pooled conditional risk and complete coverage, averaging five fixed fits. They are descriptive curves, not uncertainty bands. Separate y-ranges keep augmented-regime risks readable.

Weak-cue corruption notably increases Coupled's accepted error, demonstrating that the deliberately surviving cue is not an innocuous convenience. Changing duration also changes the temporal expo-

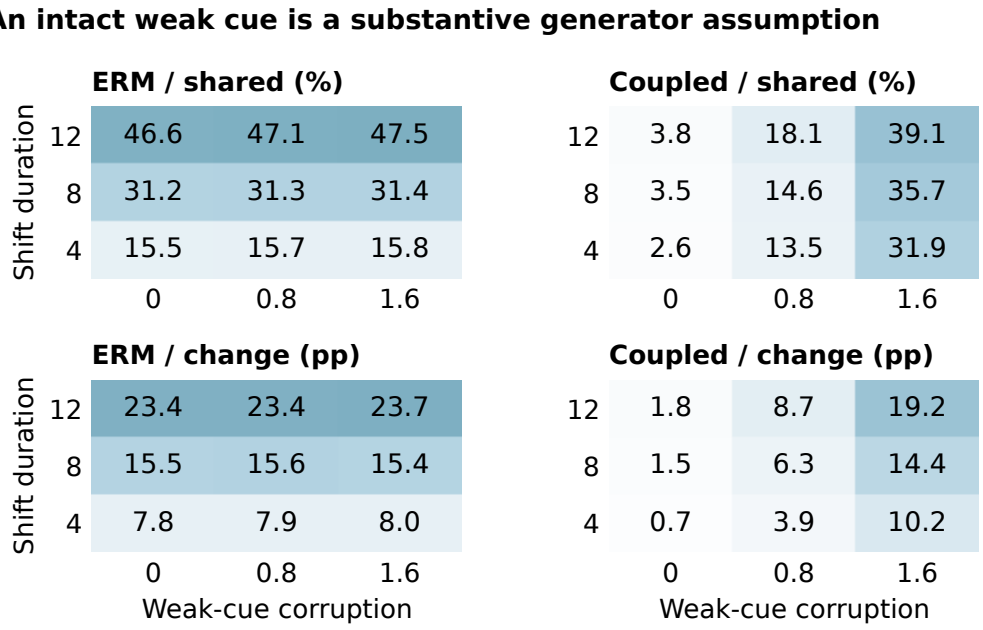


Figure 12: Retained generator-assumption stress test. Executed primary-fit results with weak-cue strength 0.8, radar amplitude multiplier 1, and temporal correlation 0.65. Rows vary strong-shift duration in frames; columns vary weak-cue corruption. Top: shared accepted error; bottom: paired change. The source summaries are reused unchanged and have no added uncertainty claim. These are feature-space stressors, not optical or RF measurements.

sure within the twelve-frame window. These existing stress runs use frozen models and thresholds, rather than retraining to each severity. Full scenario inventories and block-level arrays remain in the historical robustness directory; only the explicitly indexed slice is reproduced in Figure 12. The stress panel does not resolve the Independent/Coupled training-dose confound.

H Artifact, photographs, and figure correspondence

The current numerical evidence is in `results/v6/`. Its manifest records model paths, test conditions, probe splits, thresholds, estimands, bootstrap conventions, numerical parity, and the Bayes calculation. `predictions.npz` contains all 56 models' current probabilities and four condition inputs. `block_arrays.npz` contains endpoint, probe, restoration, eight-cell gate, joint-state, interaction, and risk–coverage arrays. `probe_heads.npz` retains fitted standardization and linear-readout parameters; clean-trained transfer is recomputed from these same heads. Current tables are generated from current CSVs rather than transcribed from images.

The new 36 checkpoints are under `models/v6_realizations/`, with training traces and selected epochs in `results/v6_training/`. The 20 primary checkpoints remain in `models/`. Earlier results, including oracle probes, NumPy auxiliary fits, original prediction archives, and physical aggregate arrays, remain separately labeled and are not canonical V6 results. No original physical victim model, synchronized raw recording, material-response library, or physical experiment has been reconstructed from those aggregates. Historical aggregate performance plots are not carried into the submission as new empirical evidence.

Figures 1 and 2 use the full photographic background objects embedded in the supplied historical qualitative PDFs, not the old overlaid detector labels, result arrows, or display insets. The face and license-plate obfuscation is already present in the supplied photograph. Only extraction and resizing are performed; no synthetic scene pixels, radar heatmaps, detections, or new physical measurements are introduced. Source filenames, embedded-object identifiers, and checksums are in the

Table 11: Current figure order and numerical provenance. Figures 1–2 are concept diagrams; all remaining figures are regenerated from explicit numerical records.

Figure	Content	Source under current results
1	Real scene / analytic plane	Supplied photograph + schematic
2	Complete-trajectory re-pairing	Supplied photograph + protocol
3	Baseline paired effects	`baselines.csv; endpoints.csv`
4	Matched / clean-fit readouts	`probes.csv; probe_transfer.csv`
5	Three restoration endpoints	`restoration.csv`
6	Distribution-aligned decoding	`probes.csv`
7	Cross-condition head transfer	`probe_transfer.csv`
8	Eight indicator-source controls	`cross_control.csv`
9	Raw / normalized interaction	`margin.csv`
10	Data-realization repetition	`replication.csv`
11	Risk–coverage curves	`risk_coverage.csv`
12	Assumption stress slice	Retained robustness `summary.csv`

asset manifest. Rights and consent documentation were not supplied and remain an author-side pre-submission requirement. The photograph is not evidence that the pictured vehicle implements the models studied here.

All newly drawn figure labels use black text; vector PDFs and editable SVGs are provided. Numerical figures have a 5.5-inch canvas matching the manuscript width and approximately 9.5–10.2-point ordinary labels, so insertion does not shrink them to unreadable sizes. The supplied ICLR2027 style and bibliography style are retained unchanged. The official guideline page was checked; a direct download of the official style archive failed, so byte identity with that remote archive is not asserted. No standalone font files are distributed.

Recompute current analyses from supplied checkpoints. From the model archive root, the canonical commands are:

```
python scripts/v6_analyze.py
python scripts/v6_controls.py
python scripts/v6_probe_transfer.py
python scripts/v6_tables.py --output paper_tables
python scripts/v6_figures.py --output paper_figures
python -m unittest discover -s tests -v
```

The V6 analysis requires the included 36 repeat checkpoints and their manifest. To generate those anew, run `pythonscripts/v6_train_realizations.py` first. The original primary training/evaluation commands are preserved separately; they are not necessary to inspect the supplied fitted-model diagnostic. The single MATLAB file reproduces Figures 3–12 in this order from embedded exported ordinates. Figures 1–2 use the supplied editable scene/diagram sources rather than numerically fabricated images. MATLAB source checks are included, but no native MATLAB execution or visual identity with the Python backend is claimed.