When Multimodality Stops Helping: Auditing Representation-Conditioned Residual Modality Utility

Anonymous authors
Paper under double-blind review

Abstract

Whether an auxiliary modality helps a prediction is not a property of the modality: it is residual to what the model's primary representation has already learned. A strong primary representation can leave a powerful auxiliary modality with nothing to add, whereas cross-modal reconstructability tempts exactly the wrong inference. We formalise this view with two estimands. *Incremental modality gain* (IMG) is the actual test-risk reduction from adding the auxiliary modality under a matched low-capacity head family, with a separate parameter-count control; *residual modality utility* (RMU) asks whether the component of the auxiliary signal that cannot be predicted from the primary representation explains the target error the primary model left. We show that cross-modal predictability implies neither conditional sufficiency nor gain. Across representation-level ladders spanning weak/generic to more task-relevant encodings on a clinical glaucoma structure–function task, CMU-MOSI, and AV-MNIST, IMG concentrates at weaker or less adapted levels and collapses toward zero at stronger, more adapted levels, while RMU tracks realised gain on every domain (seed-cluster bootstrap Spearman $\rho = 0.651$, 95% CI $[0.483, 0.825]$ pooled; per-domain CIs exclude zero on both public benchmarks) and survives negative, missingness, and capacity controls. An independent sweep over seven clinical backbones shows the same pattern. RMU is a capacity-conditional diagnostic, not conditional mutual information. These results support representation-conditioned residual utility as an operational diagnostic, not a universal sufficiency test.

1 Introduction

When does an auxiliary modality carry information that a prediction model can actually use? The textbook answer is conditional: A helps when $Y \not\perp\!\!\!\perp A \mid Z$, where Z is what the model already knows. In applied multimodal learning this conditional question is routinely replaced by two easier ones—*does a fused model score better?* and *can one modality be reconstructed from the other?*—and both substitutions are wrong in the cases that matter.

A fused model can improve because its head has more parameters rather than more information, and reconstructability measures redundancy of the observed signals rather than target-relevant residual information. Hence the central claim of this paper: *modality utility is not intrinsic to a modality; it is residual to what the current primary representation has already learned.* Adding a modality to a weak encoder can yield large gains that vanish for a stronger encoder without any change in the data distribution; conversely, a scientifically important auxiliary modality can look useless in a fusion benchmark that does not establish its information as redundant. The practical question is then not *is this modality informative?* but *which representations make collecting it worthwhile?*.

Reasoning about modality value therefore requires holding the data distribution fixed while varying the representation, and separating the gain realised on held-out data from raw signal redundancy and from target-relevant utility left in the auxiliary signal—IMG, RS,

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.

and RMU. We give a construction in which predictability approaches one while gain stays strictly positive: the residual carries the only target-relevant signal.

We evaluate on three deliberately different fixed-distribution ladders: clinical glaucoma (fundus → visual field, OCT auxiliary), CMU-MOSI (text → sentiment, acoustic auxiliary), and AV-MNIST (image → digit, audio-spectrogram auxiliary), each with four primary representation levels; the public benchmarks are performance-ordered weak to strong, whereas the glaucoma ladder is ordered by task adaptation and spatial structure. Gain is largest at the weaker or less adapted levels and collapses at stronger levels; RS does not track the collapse; RMU tracks realised gain on every domain under a pre-specified go/no-go statistic, is consistent with zero in the pre-specified control summaries, and has the strongest pooled association when compared against six pre-specified baseline diagnostics (a PLS-2 probe is numerically stronger in a separate probe-family sensitivity). An independent seven-backbone XGBoost sweep with exact Shapley values reproduces the pattern with a different estimand. The clinical evidence is retrospective, without external or cross-device validation, and never supports deployment claims.

Our contributions are:

1. **Representation-conditional formulation of modality value.** IMG, RS, and RMU separate three notions that "the modalities encode the same information" conflates, and we show that high cross-modal predictability does not imply conditional sufficiency (§3).
2. **Matched, cross-fitted estimation protocol.** One ridge head class, validation-only penalty selection, out-of-fold nuisance residuals, classification and regression variants, and pre-specified negative, missingness, and capacity controls (§4).
3. **Three-domain evidence with independent clinical replication.** Representation-level ladders on a clinical task and two public benchmarks, a pre-specified gate, a head-to-head comparison against six alternative diagnostics, and an independent seven-backbone sweep under a separate contract (§5–§8).

2 Related work

Fusion gains and modality imbalance. Multimodal benchmarks consistently find strong unimodal baselines and severe modality imbalance (Baltrušaitis et al., 2019; Liang et al., 2021); fusion architectures and balancing objectives improve performance by redistributing optimisation across modalities (Tsai et al., 2019; Pham et al., 2019; Wang et al., 2020; Peng et al., 2022; Li et al., 2023). Modality importance in this literature is measured by leave-one-out retraining, test-time permutation, or fusion attention—each quantifies how much a given trained fusion model relies on an input. We instead hold the data fixed and vary the representation, turning "does the modality help?" into "under what representation does it help?".

Interaction and contribution attribution. EMAP projects a trained model's predictions onto additive unimodal structure to decide whether interactions helped that model (Hessel & Lee, 2020); MM-SHAP attaches Shapley usage proportions to modalities (Parcalabescu & Frank, 2023); a recent inference-time analysis decomposes internal embeddings by partial information decomposition (Amit et al., 2025). These methods audit one model with its prediction function in hand; none is conditioned on representation or validated against realised held-out gain. RMU instead operates on frozen splits without the fusion model and is falsifiable against IMG.

Residualisation, conditional independence, and probing. Residualisation is the device behind partial regression and the Frisch–Waugh–Lovell theorem (Hastie et al., 2009); cross-fitting keeps nuisance estimation out of structural estimates in double/debiased machine learning (Chernozhukov et al., 2018). Conditional-independence testing is hard in high dimension (Shah & Peters, 2020), and HSIC/MI measures (Gretton et al., 2005) have power limits at our sample sizes; we do not test independence. Probes confound representation

quality with probe capacity (Hewitt & Liang, 2019; Belinkov, 2022); we import that control discipline into *modality* evaluation.

Ophthalmic multimodal prediction. Fundus and OCT backbones already reach high accuracy for structural glaucoma labels (Christopher et al., 2021); predicting perimetric sensitivity—a functional, spatially structured target—from learned fundus representations is a different question, and controlled estimates of what OCT adds beyond such representations remain rare. The clinical task is a case study here, not a source of deployment claims.

3 Gain, predictability, and residual utility

Let $Z \in \mathbb{R}^{d_Z}$ be a frozen primary representation, $A \in \mathbb{R}^{d_A}$ an auxiliary modality, and $Y \in \mathcal{Y}$ the target. Figure 1 summarises the fixed-data setup and the three distinct quantities defined below. Fix a head class $\mathcal{G}$ for $g(Z)$, a fusion class $\mathcal{M}$ for $m(Z, A)$, and a proper loss ℓ (squared error; Brier score (Brier, 1950) for classification). With penalties selected on validation, the **incremental modality gain** is

$$\mathrm{IMG} \;=\; \widehat{R}(\mathcal{G}) - \widehat{R}(\mathcal{M}), \qquad \widehat{R}(\cdot) = \tfrac{1}{n_t} \textstyle\sum_{i \in \mathrm{test}} \ell(y_i, \hat{y}_i). \tag{1}$$

Cross-modal predictability is the held-out reducibility of the auxiliary signal under the same low-capacity head class,

$$\mathrm{RS} \;=\; 1 - \frac{\sum_i \|a_i - \hat{h}(z_i)\|^2}{\sum_i \|a_i - \bar{a}\|^2}. \tag{2}$$

RS measures redundancy of the observed signal. It is deliberately not called representation sufficiency.

Proposition 1. *If $Y \perp A \mid Z$, the population Bayes risk using (Z, A) equals the population Bayes risk using Z alone, for squared loss in regression and for any strictly proper scoring rule in classification.*

This is an unrestricted population statement over the Bayes-optimal decision classes (proof Appendix A); it does not imply equality of finite-sample risks under restricted classes—the operational IMG is head-class conditional, hence the matched-head protocol and the parameter-count control. The proposition licenses dropping A only under conditional independence, and high RS does not imply that condition. Consider

$$A = Z + \varepsilon_A, \qquad Y = f(Z) + \beta\, \varepsilon_A + \varepsilon_Y, \tag{3}$$

with independent mean-zero errors and $\beta \neq 0$: RS approaches 1 as the predictable component dominates, yet $Y \not\perp\!\!\!\perp A \mid Z$ through the residual ε_A, and optimal IMG is strictly positive. RS measures the predictable part of the signal; IMG depends on the residual's relation to the target. A stronger Z can absorb the predictable part without the residual pathway—how RS can rise while gain disappears.

Residual modality utility. With $g^\star(z) = \mathbb{E}[Y \mid z]$ and $h^\star(z) = \mathbb{E}[A \mid z]$ (or class-probability and conditional-feature means for classification), form residuals $r_Y = Y - g^\star(Z)$ and $r_A = A - h^\star(Z)$. RMU is the risk reduction from a low-capacity probe q of r_Y on r_A:

$$\mathrm{RMU} \;=\; \mathbb{E}[\ell(r_Y, 0)] - \mathbb{E}[\ell(r_Y, q(r_A))]. \tag{4}$$

Population RMU is zero under conditional independence and positive in model equation 3 whenever $\beta \neq 0$. Three notions stay separate—signal redundancy (high RS), target-relevant redundancy, and conditional independence; the first implies neither other. RMU is not conditional mutual information: bounded by the chosen head classes and finite-sample error, a positive value diagnoses realised residual target information, not prospective gain.

Why the head class enters the estimand. IMG depends on the fusion head's capacity as much as on its access to A, and RMU on what the probe can express; matching head classes across cells is therefore part of the definition, not an implementation convenience. A gain measured with a larger or differently tuned head is a joint statement about information and capacity.

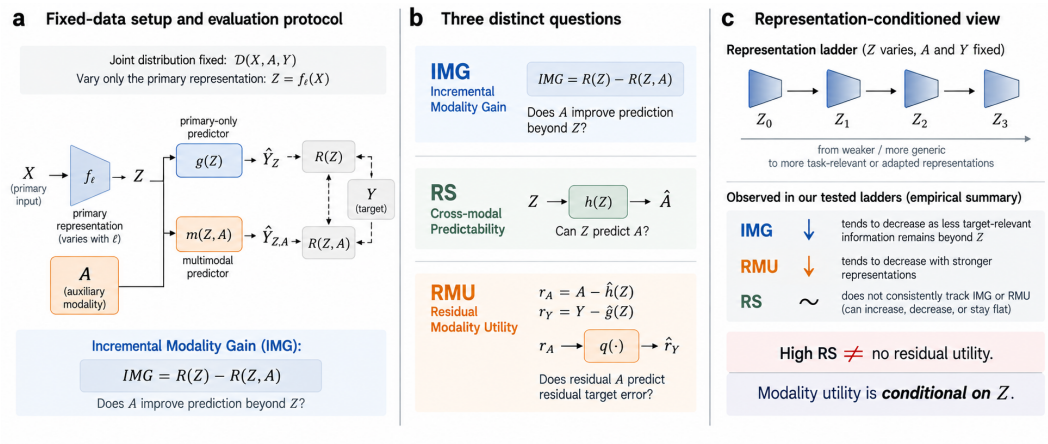

Figure 1: Representation-conditioned modality utility. (a) Under a fixed joint distribution over primary input X, auxiliary modality A, and target Y, only the learned primary representation Z varies; primary-only and auxiliary-augmented predictors define realised incremental gain. (b) IMG measures realised gain, RS predictability of A from Z, and RMU tests whether the component of A unexplained by Z predicts residual target error. (c) Across these ladders IMG and RMU shrink at more task-relevant levels whereas RS need not track the trend; high cross-modal predictability therefore does not establish that residual target-relevant utility is absent.

4 Estimation protocol

Matched low-capacity heads. Every head—base, fusion, cross-modal, and probe—is a ridge regression (multitask ridge for vector targets and one-hot classes) with an unpenalised intercept and standardised inputs. Sharing one head class by construction means IMG and RMU are never compared across heterogeneous estimators. Penalties come from a fixed grid against validation loss only.

Validation discipline. Architectures, checkpoints, and penalties are chosen without test access from frozen YAML configs and fixed validation grids; the glaucoma spatial variant is validation-selected. Each split is evaluated once per cell, cached test metrics never feed later decisions, and all such accesses are reconstructable from the registry.

Cross-fitted residuals. The nuisance heads $\hat{g}, \hat{h}$ are $K = 5$ out-of-fold fits (fixed fold seeds), so residuals never come from in-sample predictions; the probe is one fixed-penalty ridge on training residuals evaluated on test residuals. For classification, $r_Y = \text{onehot}(Y) - \hat{p}_g(Z)$, $r_A = A - \hat{h}(Z)$, and the Brier score of the simplex-projected $\hat{p}_g + q(r_A)$ is compared with that of $\hat{p}_g$.

Ladders and controls. Each dataset supplies four frozen primary representations ordered by increasing task adaptation (weak to strong in base performance on the public benchmarks), with A, Y, splits, grids, and seeds fixed; nothing else may differ between cells. Controls are: within-split *shuffle* of A; *train-only shuffle*; *zero* input (RS reported NaN, not forced to zero); *missingness* at 25/50/75%; and a *capacity* control in which the base head receives seeded Gaussian inputs of width d_A, matching the fusion head's input-parameter count with no information. Grids, dimensions, fold seeds, and numerical details are Appendix B.

Sign and reporting rule. IMG and RMU are signed and reported without truncation: a negative cell is kept, and RS stays undefined at zero variance. No cell, seed, or control is excluded post hoc; the only predeclared level exclusion (AV-MNIST A0) is always shown as a sensitivity.

Dataset	Level	Base risk	IMG	RS	RMU
Glaucoma†	G0	46.79	7.87 ± 1.29	0.351	6.89 ± 1.22
	G1	26.96	0.039 ± 0.013	0.572	0.0070 ± 0.104
	G2	27.22	0.0041 ± 0.0028	0.663	-0.044 ± 0.032
	G3	27.04	0.0030 ± 0.0031	0.669	-0.053 ± 0.034
AV-MNIST‡	A0	1.17	0.023 ± 0.0000	-0.482	0.244 ± 0.0000
	A1	0.611	0.022 ± 0.0071	0.0006	-0.0003 ± 0.0099
	A2	0.478	0.0035 ± 0.0001	0.0025	-0.0071 ± 0.0002
	A3	0.456	0.0036 ± 0.0001	0.0028	-0.0049 ± 0.0001
CMU-MOSI‡	T0	2.61	0.072 ± 0.0086	-0.034	0.080 ± 0.012
	T1	1.83	0.027 ± 0.0000	-0.033	0.0071 ± 0.0018
	T2	1.80	0.0028 ± 0.0011	-0.026	0.0023 ± 0.0022
	T3	1.67	0.0015 ± 0.0007	-0.029	0.0008 ± 0.0014

†dB^2; ‡Brier / MSE as labelled. Seed-mean summaries; per-seed cells in Appendix D. MOSI RS is near zero because pooled audio is not predictable from text; RMU still detects the weak positive gain.

Table 1: Real-condition ladder cells. IMG and RMU are both large at the weaker or less adapted levels and collapse at stronger, more adapted levels; RS shows no such collapse and can move opposite to gain.

5 Experiments: three fixed-distribution ladders

Settings. *Glaucoma* (retrospective clinical cohort): patient-group split 1,125/208/222 records with committed manifest hashes and no patient overlap; Y is 59 visual-field sensitivities (dB), A six OCT measures. G0 is frozen ImageNet (Russakovsky et al., 2015) pooling through a seeded 32-d Gaussian bottleneck; G1 the pooled token of a task-trained ConvNeXt-Tiny (Liu et al., 2022); G2 its 12,544-d spatial tokens; G3 a validation-selected spatial variant; 3 seeds. *AV-MNIST*: MultiBench 55k/5k/10k split, image primary, flattened 112×112 audio spectrogram, digit target. A0 raw pixels; A1 a frozen random 64-d projection; A2/A3 learned MLPs (128/256-d); final checkpoints, 5 seeds. *CMU-MOSI*: canonical 1,284/229/686 split, GloVe-300 text primary, 5-d COVAREP audio, sentiment in $[-3, 3]$. T0 a 16-d Gaussian bottleneck; T1 masked-mean GloVe; T2 a GRU-64 (Cho et al., 2014) output mean; T3 a bidirectional GRU-64; fixed 200-epoch schedule. The three settings are deliberately heterogeneous in target type, dimensionality, and auxiliary structure.

Gain concentrates at weaker or less adapted levels. On CMU-MOSI base MSE improves 2.61/1.83/1.80/1.67 while audio IMG decays $0.072 \to 0.027 \to 0.003 \to 0.002$; on glaucoma, OCT IMG averages 7.87 dB^2 at G0 but is ≤ 0.05 at G1 and ≤ 0.005 at G2/G3; on AV-MNIST, audio IMG is 0.023/0.022/0.004/0.004 while base accuracy rises 0.09/0.56/0.64/0.65. On CMU-MOSI and AV-MNIST the auxiliary helps most at the weakest representations and becomes very small at the strongest; on glaucoma the largest swing occurs between G0 and G1, where OCT IMG falls from ≈7.9 to ≈0.04 dB^2 (Fig. 2, Table 1). Explicit caveat: under the linear head G1/G2/G3 base risk is not monotone— glaucoma geometry varies more than probe performance, so G0–G3 is a task-adaptation ladder, not a performance-ordered strength ladder.

Pre-specified gate. We pre-specified the following evaluation gate for the main rank-tracking claim: RMU tracks IMG on a public dataset only if the seed-cluster bootstrap Spearman $\rho(\text{RMU}, \text{IMG})$, $B = 10{,}000$, is positive with 95% CI excluding zero; the paper-level claim requires both public datasets. Results: AV-MNIST $\rho = 0.767$ $[0.599, 0.876]$; CMU-MOSI 0.974 $[0.919, 0.992]$; glaucoma 0.845 $[0.333, 1.00]$ (3 seeds, wide interval); pooled across 13 dataset×seed clusters 0.651 $[0.483, 0.825]$ (Fig. 3). The gate passes on both benchmarks. Its scope is narrow: rank tracking within tested levels, not magnitude equality or transfer beyond the frozen heads and datasets. In the reverse direction $\rho(\text{RS}, \text{IMG})$ is -0.965 on glaucoma and -0.201 (CI $[-0.436, -0.096]$) pooled: RS is negatively associated with IMG in these tested cells. Excluding the miscalibrated A0 level the AV gate stays pos-

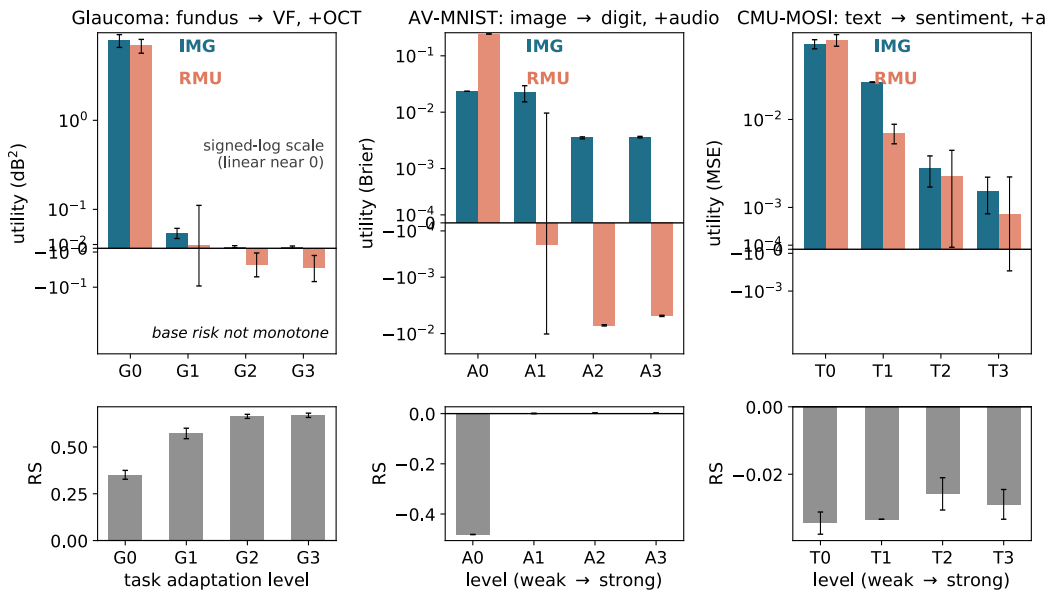

Figure 2: Representation-level ladders (seed means ± SD). Top: IMG and RMU on signed-log scales—both large at weak levels, small at more adapted levels on all three domains. Bottom: RS does not track the collapse (and is large at strong glaucoma levels). Glaucoma is ordered by task adaptation/spatial structure, not performance (base risk is not monotone); the public benchmarks are weak-to-strong.

itive (0.571 [0.150,0.883]) as does the pooled gate (0.637 [0.384,0.849]). Secondary analyses (mixed model; all intervals, no p-values) are in Appendix G.

Reading the magnitudes. At weak levels RMU is comparable in scale to realised IMG (glaucoma G0 6.89 vs 7.87 dB^2; MOSI T0 0.080 vs 0.072); AV-MNIST A0 is the exception (RMU 0.244 vs IMG 0.023), for the calibration reason given in §6. At strong levels both are essentially zero, and RMU's small negative values at glaucoma G2/G3 mean no residual signal is extracted, not that OCT harms prediction. The glaucoma swing after task adaptation is consistent with representation-dependent redundancy under this protocol—not evidence that OCT information has been captured or replaced; on MOSI the gain concentrates at T0/T1 and its ordering comes from the residual-probe contrast rather than redundancy; on AV-MNIST, audio gain is approximately 0.022 at A1 and falls to ≤ 0.005 Brier at A2/A3.

6 Controls and alternative explanations

At the predeclared control summaries, seed-mean shuffled effects collapse toward zero or below (glaucoma IMG 0.038, RMU −0.100; MOSI IMG −0.0022, RMU −0.0062); zeroing A gives the same aggregate pattern at calibrated levels. Capacity controls show no systematic positive RMU in the aggregate summaries (−0.416 glaucoma; −0.0082 MOSI). At the predeclared missingness summaries (all-level means dominated by G0/T0), increasing missingness (25/50/75%) attenuates IMG and RMU; the near-zero non-weak cells are not monotone. On MOSI the capacity control shows a tiny chance *IMG* (0.013) but no RMU—the discrimination the probe adds. Individual-seed fluctuations, including occasional positive seeds, are retained in Appendix E.

Calibration caveat. At AV-MNIST A0 the ridge base is worse than uniform on test (Brier 1.17 vs 0.90). RMU there (0.244) partly repairs base calibration, and shuffled/zeroed inputs are correspondingly elevated (0.050/0.019). At calibrated levels (A1–A3 and both

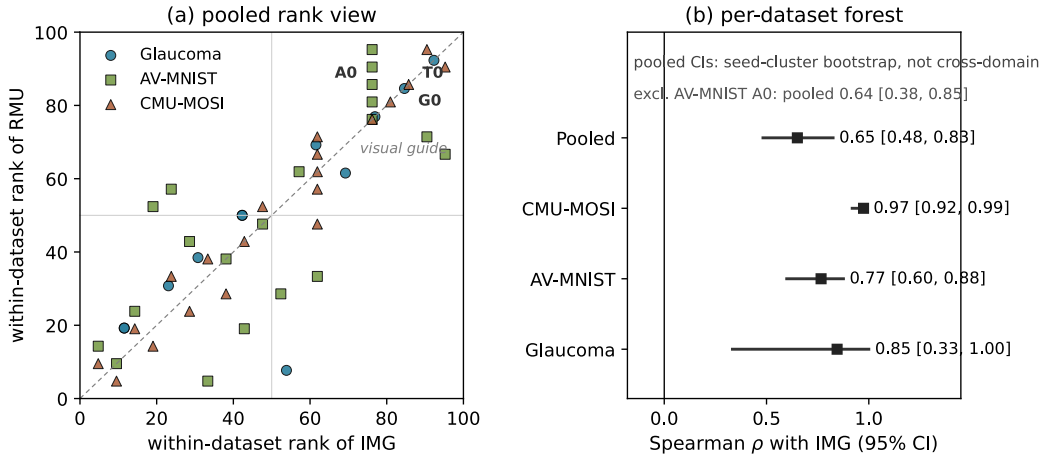

Figure 3: The rank-tracking claim. (a) Within-dataset ranks of RMU vs IMG (dataset-specific shapes; weakest labels shown). (b) Per-dataset and pooled Spearman ρ with seed-cluster bootstrap 95% CIs; public and pooled CIs exclude zero. The top annotation reports the A0-exclusion sensitivity (Table 2).

other domains) controls are consistent with zero in seed-mean summaries. We read RMU's *ranks* at A0, not its magnitude; the gate is unchanged without A0.

Synthetic identification. Fixed-weight data-generating processes confirm high RS with zero IMG and zero RMU under conditional independence, high RS with positive IMG and RMU under model equation 3, and RMU monotone in residual coupling (Appendix I).

7 Alternative diagnostics

Is RMU merely residualisation renamed? Six alternatives were computed on the identical cells—same splits, nuisance heads, residuals, folds, and seeds—and scored by the same gate (Table 2; per-dataset values Appendix G). Against the six pre-specified conventional baseline diagnostics, RMU has the strongest pooled rank association with realised gain, 0.651 $[0.483, 0.825]$; PLS-1 is close, 0.620 $[0.377, 0.847]$; Gaussian-kernel HSIC reaches 0.483 (excluding AV-MNIST); partial correlation, linear HSIC, regularised CCA, and descriptive in-sample residual R^2 all rank cells markedly less reliably. Dropping every weakest level attenuates all methods: RMU remains positive at 0.369 $[0.023, 0.711]$. Of the six baseline diagnostics, Gaussian-kernel HSIC is the only one whose interval remains above zero (0.313 $[0.038, 0.674]$); the remaining baseline intervals span zero (partial correlation even turns negative). Weak representations contribute disproportionately to every gate, but RMU's advantage is not an artefact of the extreme cells; dropping G0 alone leaves RMU at 0.582 $[0.378, 0.777]$, still ahead of every alternative.

Probe-family robustness is mixed: among linear low-capacity probes PLS-2 is numerically strongest (0.713 $[0.546, 0.882]$), ridge and PLS-1 close, but fixed-budget ElasticNet fails to converge on AV-MNIST (0.033) and a small MLP is weak (0.291 $[-0.024, 0.540]$). The pattern is robust to reasonable linear probes, not every family. PLS-1 cuts both ways: one held-out latent direction captures most of the ordering, arguing against probe overfitting, but its wider interval and weak glaucoma ranking (0.31) show it lacks the cross-fitted nuisance control; RMU's advantage is the matched operational protocol, not a new residualisation principle.

Diagnostic	pooled Spearman ρ with IMG [95% CI]
RMU, cross-fitted ridge (ours)	**0.651 [0.483, 0.825]**
PLS-1 probe (held out)	0.620 [0.377, 0.847]
HSIC, Gaussian kernel	0.483 [0.355, 0.824]
Partial correlation (mean abs.)	0.419 [0.317, 0.567]
HSIC, linear kernel	0.388 [0.234, 0.544]
Regularized CCA (first canonical)	0.273 [0.161, 0.494]
Residual R^2 (in-sample; descriptive)	0.270 [0.182, 0.414]

Table 2: Head-to-head diagnostics on identical frozen cells: pooled Spearman rank correlation with actual IMG, seed-cluster bootstrap 95% CIs. RMU is strongest against the six pre-specified conventional baselines; PLS-1 is close; unsupervised dependence measures are markedly weaker. The Gaussian-HSIC pooled statistic uses only the domains on which it is computed (glaucoma and MOSI) and therefore has different support from the full three-domain pooled statistics. Probe-family sensitivities (e.g., PLS-2) are separate and shown in Appendix G.

8 Independent clinical backbone sweep

This sweep is corroboration under a different observational contract, not a second identification test. A separately executed protocol—three predeclared seeds, fixed XGBoost head, patient-group bootstrap, exact Shapley values—compared seven clinical vision representations under one matched global-pooling contract. OCT contribution is small for the partially task-adapted RETFound/DINOv2 towers (0.005–0.016 dB), larger for their frozen counterparts (≈ 0.20 and 0.34–0.36 dB), intermediate for the task-trained ConvNeXt (0.11–0.13 dB), moderate for RET-CLIP (≈ 0.17 dB), and largest for the weakest EyeCLIP tower (0.65–0.71 dB); clinical text contributes negligibly throughout (Fig. 4), with seed-mean text contributions within $[-0.059, 0.053]$ dB across the four encoders and raw/redacted conditions.

Sweep contract. The full $7 \times 4 \times \{\text{raw}, \text{redacted}\}$ matrix (56 cells) uses one XGBoost contract; modality value is the exact permutation Shapley contribution to test MAE59 with patient-group resampling. The head and estimand differ from the RMU ladders, so this is supporting evidence, never pooled (Appendix H).

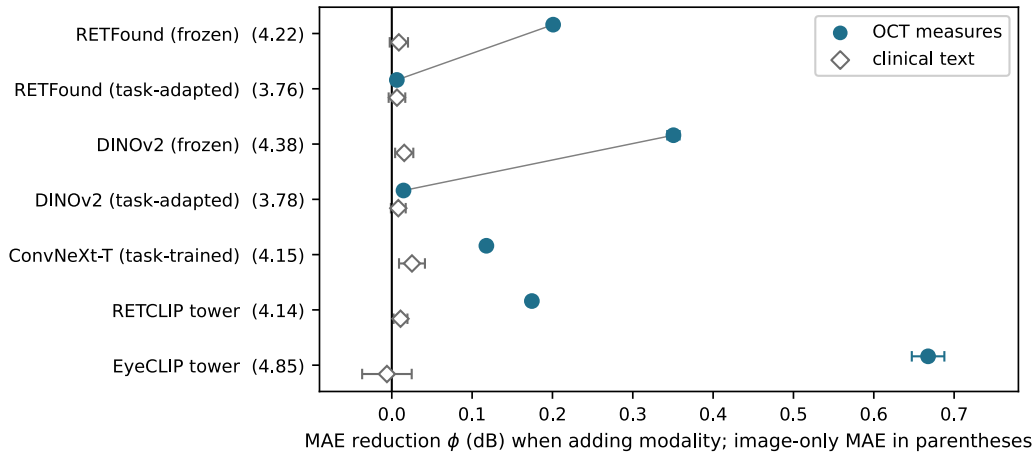

Figure 4: Independent clinical sweep under a different head/estimand; supporting evidence only (test, mean ± seed SD). OCT contributions (circles) vs clinical text (diamonds); paired frozen/adapted families connected by thin grey lines; image-only MAE in parentheses.

9 Limitations and conclusion

RMU is conditional on representation, probe capacity, and head class; it is not conditional mutual information and cannot detect structure the ridge class cannot express or the fixed grid over-regularises—the sensitivities above quantify that dependence. The glaucoma ladder is not monotone in base performance, the cohort lacks external or cross-device validation, and MOSI's pooled audio is coarse. AV-MNIST A0 magnitudes partly reflect base-calibration repair; the main rank-tracking conclusions remain under the A0-exclusion sensitivity. The paper-level gate is defined on the two public benchmarks; glaucoma is supporting evidence given its wide interval and the split's prior project use (an earlier architecture study and the backbone sweep). Nothing here is a universal law: a pre-registered cross-dataset forecast is the natural next step. $K = 5$ cross-fitting adds residual noise in exchange for removing in-sample bias, and null verdicts are specific to the two-stage protocol—a co-trained model could recover gains fixed heads leave unused. Other correlations and the mixed model are secondary analyses, so no multiplicity correction applies to the gate. Risks stay in native units; cross-domain comparisons are rank comparisons only.

Conclusion. Modality utility is residual, not intrinsic: across the tested ladders realised gain concentrates at weaker or less adapted levels, cross-modal predictability alone can point the wrong way, and the target-relevant residual component tracks whether the modality improves predictions, with an independent clinical sweep showing the same pattern. The practical question is therefore not whether an auxiliary modality is informative in isolation, but whether it contains target-relevant information that remains usable beyond the current representation.

References

Padegal Amit, Omkar Mahesh Kashyap, Namitha Rayasam, Nidhi Shekhar, and Surabhi Narayan. Quantifying modality contributions via disentangling multimodal representations. *arXiv preprint arXiv:2511.19470*, 2025.

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

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of EMNLP 2014*, pp. 1724–1734, 2014.

Mark Christopher, Akram Belghith, Robert N. Weinreb, Michael H. Goldbaum, Christopher Bowd, Marc Prydderch, Oscar A. Jimenez del Toro, Paolo Favaro, Carla Várgas, et al. Deep learning approaches to identify eyes with glaucomatous optic neuropathy from fundus photographs and optical coherence tomography. In *Ophthalmology*, volume 128, pp. 1534–1548, 2021.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep: A collaborative speech signal processing repository. *Computer Speech & Language*, 28(4): 977–993, 2014.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert–schmidt norms. *Algorithmic Learning Theory*, pp. 63–77, 2005.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. In *Springer Series in Statistics*, 2009.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877, 2020.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 2733–2743, 2019.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multimodal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zongkai Deng, Xiang Ye, Xingyu Fan, Sravan Chen, Kriti Bhoopchand, Ashley Wu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Letitia Parcalabescu and Anette Frank. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 4032–4059, 2023.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8238–8247, 2022.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6892–6899, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Rajen D. Shah and Jonas Peters. Hardness of the conditional independence problem and a generalized covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019.

Wei Wang, Du Tran, and Matt Feiszli. Optimizing multimodal fusion via re-weighting and transforming modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10222–10231, 2020.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gesture and verbal messages. In *IEEE Intelligent Systems*, volume 31, pp. 82–88, 2016.

A Proof of Proposition 1

The proof concerns the *unrestricted* population Bayes decision classes; no finite ridge class, capacity bound, or regulariser is assumed.

Regression. With squared loss the Bayes predictor given inputs is the conditional mean of Y. Under $Y \perp A \mid Z$,

$$p(y \mid z, a) = p(y \mid z) \;\Rightarrow\; \mathbb{E}[Y \mid Z = z, A = a] = \mathbb{E}[Y \mid Z = z] = g^\star(z),$$

so the Bayes-optimal decision using (Z, A) is the same function of z as the Bayes-optimal decision using Z alone, and the two population Bayes risks coincide:

$$\inf_d \mathbb{E}[(Y - d(Z, A))^2] = \inf_d \mathbb{E}[(Y - d(Z))^2].$$

Conversely, under the residual DGP $A = Z + \varepsilon_A$, $Y = f(Z) + \beta\varepsilon_A + \varepsilon_Y$, $\mathbb{E}[Y \mid Z, A] = f(Z) + \beta(A - Z) \neq f(Z)$ for $\beta \neq 0$, so the Bayes risk using (Z, A) is strictly lower.

Classification. For K classes let $p_k^\star(z) = \Pr(Y = k \mid Z = z)$. Under conditional independence $\Pr(Y = k \mid Z, A) = \Pr(Y = k \mid Z) = p_k^\star(Z)$, so the Bayes-optimal class-probability vector using (Z, A) is the same vector as that using Z alone. For any strictly proper scoring rule S, the unique expected-score minimiser is the true class-probability vector; hence the minimised population score given (Z, A) equals that given Z. The Brier score is the squared-error special case.

Why this does not force zero operational IMG. The theorem concerns unrestricted population Bayes risks. Operational IMG is computed under restricted ridge head classes with fitted coefficients, and can deviate from zero through approximation error of the Bayes decision, regularisation, numerical optimisation, and finite-sample variation. IMG is therefore an empirical, head-class conditional estimand; conditional independence does not guarantee zero finite-sample IMG for an arbitrary restricted head class.

RMU null under conditional independence. The residuals $r_Y = Y - g^\star(Z)$, $r_A = A - h^\star(Z)$ obey $\mathbb{E}[r_Y \mid Z] = \mathbb{E}[r_A \mid Z] = 0$, and under conditional independence additionally $\mathbb{E}[r_Y \mid r_A, Z] = 0$, so population RMU is zero for any probe class. Under the residual DGP, cross-fitted residuals satisfy $r_A = \varepsilon_A + o_p(1)$ and $r_Y = \beta\varepsilon_A + \varepsilon_Y + o_p(1)$; for an unrestricted linear probe the explained fraction is $\beta^2\mathrm{Var}(\varepsilon_A)/\mathrm{Var}(r_Y)$, increasing in $|\beta|$. The synthetic experiments of Appendix I verify the monotonicity at finite sample.

B Estimation details

Matched ridge heads. All heads are ridge regressions—multitask ridge for vector targets and one-hot classes—with unpenalised intercepts. Inputs are standardised with training means/scales; regression labels are standardised for fitting and converted back to the original scale for risk reporting. Penalties come from fixed grids, selected on validation loss only: regression $\alpha \in \{0.1, 1, 10, 100, 1000\}$; classification $\alpha \in \{1, 10, 100, 1000\}$. The probe penalty is fixed at 1 to avoid an additional validation choice.

Cross-fitting. Nuisance heads $\hat{g}, \hat{h}$ are out-of-fold fits with $K = 5$ folds; fold assignments are deterministic per cell from fixed fold seeds, so residuals never use in-sample predictions. The probe is one fixed-penalty ridge on training-fold residuals and is evaluated once on held-out test residuals.

Classification (option 1, frozen before results). $p_g(Z)$ is a simplex-projected K-vector from ridge on one-hot Y. The Brier residual is $r_Y = \text{onehot}(Y) - p_g(Z)$; its squared norm integrates the multiclass Brier score of g. With $r_A = A - \hat{h}(Z)$, the probe output gives $p_{\text{aux}} = \Pi_\Delta(p_g + q(r_A))$, where Π_Δ is Euclidean projection onto the probability simplex. RMU is the held-out difference $\text{Brier}(p_g) - \text{Brier}(p_{\text{aux}})$. We fixed this weaker option (rather than logit-gradient targets) before seeing any public-benchmark result.

Representation dimensions and head sizes. Glaucoma: G0 width 32 (seeded Gaussian bottleneck over ImageNet pooled features); G1 pooled token of a task-trained ConvNeXt-Tiny; G2 width 12,544 spatial tokens; G3 the validation-selected spatial variant; $d_Y = 59$. AV-MNIST: A0 width 784; A1 width 64; A2 width 128; A3 width 256; 10 classes. CMU-MOSI: T0 width 16; T1 width 300; T2 width 64; T3 width 128; $d_Y = 1$, $d_A = 5$. The base head has $(d_Z + 1)d_Y$ parameters and the fusion head $(d_Z + d_A + 1)d_Y$.

Numerical implementation. For AV-MNIST, where d_A is large, ridge Gram matrices are accumulated in float64 on GPU; predictions agree with `sklearn` Ridge (Hastie et al., 2009) to within 5×10^{-7} on shared problems. Encoder checkpoints are final checkpoints only; no best-model selection against validation accuracy is performed after training.

C Public-dataset preprocessing and licensing

AV-MNIST is loaded from the MultiBench (Liang et al., 2021) AV-MNIST pack; images are reshaped to $(1, 28, 28)$ and audio to $(1, 112, 112)$ and scaled by $1/255$ exactly as in the upstream `get_data.py` (verified against the MultiBench source), using the canonical 55k/5k/10k split. **CMU-MOSI** uses the MultiBench aligned `mosi_data.pkl` (GloVe-300 (Pennington et al., 2014) text, 5-d COVAREP-derived (Degottex et al., 2014) audio, OpenFace vision unused, sentiment in $[-3, 3]$ following the original protocol (Zadeh et al., 2016)) with the canonical 1,284/229/686 split; segment and video identifiers have zero cross-split overlap. Both packs are research-use resources that we do not redistribute: raw features remain outside version control; loader code, split sizes, source URLs, and file hashes are committed. Clinical images and tabular records never leave the controlled clinical storage.

D All cell results

Table 3 lists every seed for every real-condition ladder cell, generated from the committed aggregate CSVs; the values match Table 1 by seed means. Control cells (train-only shuffle, missingness, capacity) are in Appendix E.

Table 3: Per-seed ladder cells (all conditions). Base, IMG and RMU are in dataset risk units (dB^2 MSE, Brier, or raw-scale MSE); '–' marks undefined RS for constant auxiliary.

Dataset	Level	Seed	Cond.	Base	IMG	RS	RMU
glaucoma	G0	13	real	48.61	8.87	0.371	7.88
glaucoma	G0	42	real	47.55	8.33	0.358	7.25
glaucoma	G0	73	real	44.21	6.42	0.325	5.53
glaucoma	G1	13	real	27.18	0.024	0.604	-0.098
glaucoma	G1	42	real	26.90	0.050	0.560	0.0090
glaucoma	G1	73	real	26.81	0.044	0.552	0.110
glaucoma	G2	13	real	27.54	0.0065	0.656	-0.014
glaucoma	G2	42	real	27.74	0.0048	0.658	-0.041
glaucoma	G2	73	real	26.37	0.0010	0.676	-0.077
glaucoma	G3	13	real	27.54	0.0065	0.656	-0.014
glaucoma	G3	42	real	27.21	0.0014	0.675	-0.067
glaucoma	G3	73	real	26.37	0.0010	0.676	-0.077
avmnist	A2	13	real	0.479	0.0036	0.0025	-0.0074
avmnist	A1	13	real	0.567	0.018	0.0005	-0.0034
avmnist	A2	13	shuffle	0.479	-0.012	-0.0055	-0.018
avmnist	A1	13	shuffle	0.567	-0.0089	-0.0057	-0.018
avmnist	A1	13	zero	0.567	0.0000	–	-0.0002
avmnist	A2	13	zero	0.479	0.0000	–	-0.0004
avmnist	A1	42	real	0.544	0.014	0.0017	-0.0062
avmnist	A2	42	real	0.478	0.0034	0.0026	-0.0072

Dataset	Level	Seed	Cond.	Base	IMG	RS	RMU
avmnist	A1	42	shuffle	0.544	-0.0091	-0.0055	-0.018
avmnist	A2	42	shuffle	0.478	-0.011	-0.0054	-0.017
avmnist	A1	42	zero	0.544	0.0000	–	-0.0002
avmnist	A2	42	zero	0.478	0.0000	–	-0.0004
avmnist	A1	73	real	0.630	0.030	0.0001	-0.0031
avmnist	A2	73	real	0.478	0.0034	0.0024	-0.0072
avmnist	A1	73	shuffle	0.630	-0.0055	-0.0058	-0.015
avmnist	A2	73	shuffle	0.478	-0.011	-0.0054	-0.017
avmnist	A1	73	zero	0.630	0.0000	–	0.0001
avmnist	A2	73	zero	0.478	0.0000	–	-0.0004
avmnist	A1	101	real	0.716	0.030	-0.0011	0.017
avmnist	A2	101	real	0.477	0.0034	0.0025	-0.0069
avmnist	A1	101	shuffle	0.716	-0.0048	-0.0061	-0.0079
avmnist	A2	101	shuffle	0.477	-0.011	-0.0054	-0.016
avmnist	A1	101	zero	0.716	0.0000	–	0.0004
avmnist	A2	101	zero	0.477	0.0000	–	-0.0004
avmnist	A1	137	real	0.597	0.020	0.0018	-0.0058
avmnist	A1	137	shuffle	0.597	-0.0051	-0.0055	-0.016
avmnist	A2	137	real	0.476	0.0037	0.0026	-0.0071
avmnist	A1	137	zero	0.597	0.0000	–	-0.0007
avmnist	A2	137	shuffle	0.476	-0.011	-0.0055	-0.016
avmnist	A0	13	real	1.17	0.023	-0.482	0.244
avmnist	A2	137	zero	0.476	0.0000	–	-0.0004
avmnist	A0	13	shuffle	1.17	-0.016	-0.021	0.048
avmnist	A3	13	real	0.456	0.0038	0.0030	-0.0048
avmnist	A0	13	zero	1.17	0.0000	–	0.019
avmnist	A3	13	shuffle	0.456	-0.011	-0.0058	-0.016
avmnist	A0	42	real	1.17	0.023	-0.482	0.244
avmnist	A3	13	zero	0.456	0.0000	–	-0.0002
avmnist	A0	42	shuffle	1.17	-0.015	-0.014	0.051
avmnist	A3	42	real	0.456	0.0035	0.0028	-0.0048
avmnist	A0	42	zero	1.17	0.0000	–	0.019
avmnist	A3	42	shuffle	0.456	-0.011	-0.0059	-0.015
avmnist	A3	42	zero	0.456	0.0000	–	-0.0001
avmnist	A0	73	real	1.17	0.023	-0.482	0.244
avmnist	A0	73	shuffle	1.17	-0.015	-0.011	0.051
avmnist	A3	73	real	0.456	0.0037	0.0027	-0.0051
avmnist	A3	73	shuffle	0.456	-0.010	-0.0061	-0.014
avmnist	A0	73	zero	1.17	0.0000	–	0.019
avmnist	A3	73	zero	0.456	0.0000	–	-0.0001
avmnist	A0	101	real	1.17	0.023	-0.482	0.244
avmnist	A3	101	real	0.455	0.0036	0.0030	-0.0048
avmnist	A0	101	shuffle	1.17	-0.016	-0.010	0.049
avmnist	A3	101	shuffle	0.455	-0.0098	-0.0059	-0.014
avmnist	A0	101	zero	1.17	0.0000	–	0.019
avmnist	A3	101	zero	0.455	0.0000	–	-0.0001
avmnist	A0	137	real	1.17	0.023	-0.482	0.244
avmnist	A3	137	real	0.455	0.0036	0.0027	-0.0049
avmnist	A0	137	shuffle	1.17	-0.016	-0.012	0.051
avmnist	A3	137	shuffle	0.455	-0.0100	-0.0060	-0.014
avmnist	A0	137	zero	1.17	0.0000	–	0.019
avmnist	A3	137	zero	0.455	0.0000	–	-0.0001
mosi	T0	13	real	2.54	0.081	-0.034	0.085
mosi	T0	13	shuffle	2.54	-0.0007	-0.049	-0.0087
mosi	T0	42	real	2.61	0.072	-0.040	0.080
mosi	T0	42	shuffle	2.61	-0.0077	-0.049	-0.034
mosi	T0	73	real	2.53	0.061	-0.032	0.065
mosi	T0	73	shuffle	2.53	-0.0066	-0.048	-0.013
mosi	T0	101	real	2.67	0.066	-0.032	0.074
mosi	T0	101	shuffle	2.67	-0.0080	-0.049	-0.022
mosi	T0	137	real	2.70	0.080	-0.035	0.097
mosi	T0	137	shuffle	2.70	0.0091	-0.049	0.011
mosi	T1	13	real	1.83	0.027	-0.033	0.0074
mosi	T1	13	shuffle	1.83	0.0010	-0.082	0.0014
mosi	T1	42	real	1.83	0.027	-0.033	0.0077
mosi	T1	42	shuffle	1.83	-0.0080	-0.101	-0.028
mosi	T1	73	real	1.83	0.027	-0.033	0.0074
mosi	T1	73	shuffle	1.83	-0.0077	-0.086	-0.018
mosi	T1	101	real	1.83	0.027	-0.033	0.0088
mosi	T1	101	shuffle	1.83	0.0006	-0.079	0.0011
mosi	T1	137	real	1.83	0.027	-0.033	0.0041
mosi	T1	137	shuffle	1.83	-0.0013	-0.082	-0.0085
mosi	T2	13	real	1.76	0.0021	-0.023	0.0011
mosi	T2	13	shuffle	1.76	0.0002	-0.050	0.0007
mosi	T2	42	real	1.80	0.0033	-0.033	0.0031
mosi	T2	42	shuffle	1.80	0.0000	-0.055	0.0004
mosi	T2	73	real	1.83	0.0026	-0.023	0.0018
mosi	T2	73	shuffle	1.83	0.0007	-0.056	0.0015
mosi	T2	101	real	1.81	0.0015	-0.028	-0.0003

Dataset	Level	Seed	Cond.	Base	IMG	RS	RMU
mosi	T2	101	shuffle	1.81	0.0006	-0.049	0.0013
mosi	T2	137	real	1.82	0.0043	-0.021	0.0056
mosi	T2	137	shuffle	1.82	-0.0005	-0.048	-0.0013
mosi	T3	13	real	1.70	0.0012	-0.032	0.0008
mosi	T3	13	shuffle	1.70	0.0004	-0.064	0.0010
mosi	T3	42	real	1.67	0.0008	-0.029	-0.0004
mosi	T3	42	shuffle	1.67	0.0001	-0.065	-0.0001
mosi	T3	73	real	1.65	0.0019	-0.027	0.0019
mosi	T3	73	shuffle	1.65	-0.0001	-0.057	-0.0003
mosi	T3	101	real	1.60	0.0012	-0.035	-0.0006
mosi	T3	101	shuffle	1.60	-0.0000	-0.059	0.0003
mosi	T3	137	real	1.70	0.0025	-0.023	0.0025
mosi	T3	137	shuffle	1.70	-0.0001	-0.057	-0.0005
mosi	T0	13	zero	2.54	0.0000	–	-0.0031
mosi	T0	42	zero	2.61	0.0000	–	-0.0037
mosi	T0	73	zero	2.53	0.0000	–	-0.0009
mosi	T0	101	zero	2.67	0.0000	–	-0.0011
mosi	T0	137	zero	2.70	0.0000	–	-0.0008
mosi	T1	13	zero	1.83	-0.0000	–	-0.0009
mosi	T1	42	zero	1.83	-0.0000	–	-0.0004
mosi	T1	73	zero	1.83	-0.0000	–	-0.0025
mosi	T1	101	zero	1.83	-0.0000	–	0.0001
mosi	T1	137	zero	1.83	-0.0000	–	-0.0009
mosi	T2	13	zero	1.76	0.0000	–	-0.0000
mosi	T2	42	zero	1.80	-0.0000	–	0.0001
mosi	T2	73	zero	1.83	-0.0000	–	-0.0001
mosi	T2	101	zero	1.81	-0.0000	–	-0.0001
mosi	T2	137	zero	1.82	-0.0000	–	0.0002
mosi	T3	13	zero	1.70	-0.0000	–	-0.0001
mosi	T3	42	zero	1.67	0.0000	–	-0.0001
mosi	T3	73	zero	1.65	0.0000	–	0.0000
mosi	T3	101	zero	1.60	-0.0000	–	0.0001
mosi	T3	137	zero	1.70	0.0000	–	0.0001

E Full controls

The table at the end of this section reports all negative, missingness, and capacity cells generated from `iclr_controls.csv`. At calibrated levels, within-split shuffle and zero conditions give IMG and RMU consistent with zero in the seed-mean summaries; individual-seed fluctuations are shown in the table. Train-only shuffling collapses gains on glaucoma (IMG 0.038, RMU −0.100), MOSI (IMG −0.0022, RMU −0.0062), and AV-MNIST A3 (IMG −0.010, RMU −0.015). At the predeclared all-level summaries, missingness at 25/50/75% attenuates both metrics on glaucoma (IMG 1.14/ 0.694/0.282 dB^2; RMU 0.988/ 0.608/0.233) and MOSI (IMG 0.017/0.010/0.0023; RMU 0.026/0.016/0.0038); the near-zero non-weak cells need not be monotone. At AV-MNIST A3 both metrics are consistent with zero or below at every rate. Capacity controls show no systematic positive RMU in the aggregate summaries (−0.416 glaucoma; −0.0082 MOSI; −0.080 A3), although individual seeds fluctuate around zero.

The AV-MNIST A0 exceptions are the calibration artefact described in the main text: the ridge base is worse than uniform on test (Brier 1.17 vs 0.90), shuffled/zeroed inputs are elevated but far below real audio (shuffle IMG −0.016, RMU 0.052), RMU stays positive under missingness even when IMG does not (0.208/0.156/0.090), and the capacity control's RMU is 0.123 for the same reason. No A0 per-cell RMU magnitude is interpreted as information content; only its ranks are used.

Full control cells (IMG/RS/RMU), generated from `iclr_controls.csv`*. "–" marks numerically degenerate RS (the AV-MNIST 12,544-dimensional random capacity control; its floating-point RS falls outside* $[-1, 1]$*); these cells enter no rank analysis.*

Dataset	Level	Seed	Control	Setting	IMG	RS	RMU
glaucoma	G0	13	capacity	random_d6	-0.088	0.836	-0.161
glaucoma	G0	13	missingness	0.25	5.15	0.033	4.81
glaucoma	G0	13	missingness	0.50	2.87	0.0074	2.50
glaucoma	G0	13	missingness	0.75	1.29	0.0035	0.949
glaucoma	G0	13	shuffle_train	train	0.028	-0.039	-0.187

Dataset	Level	Seed	Control	Setting	IMG	RS	RMU
glaucoma	G0	42	capacity	random_d6	-0.475	0.838	-0.227
glaucoma	G0	42	missingness	0.25	4.04	0.063	3.19
glaucoma	G0	42	missingness	0.50	2.83	0.041	2.32
glaucoma	G0	42	missingness	0.75	0.599	0.0074	0.129
glaucoma	G0	42	shuffle_train	train	0.134	-0.040	-0.071
glaucoma	G0	73	capacity	random_d6	-0.835	0.836	-0.340
glaucoma	G0	73	missingness	0.25	4.57	0.035	4.54
glaucoma	G0	73	missingness	0.50	2.67	0.0028	2.78
glaucoma	G0	73	missingness	0.75	1.49	-0.039	1.96
glaucoma	G0	73	shuffle_train	train	0.418	-0.0066	-0.0098
glaucoma	G1	13	capacity	random_d6	0.043	0.970	-1.29
glaucoma	G1	13	missingness	0.25	0.0004	0.105	-0.0078
glaucoma	G1	13	missingness	0.50	-0.0080	0.031	-0.047
glaucoma	G1	13	missingness	0.75	-0.0039	0.012	0.053
glaucoma	G1	13	shuffle_train	train	-0.067	-0.026	-0.184
glaucoma	G1	42	capacity	random_d6	0.060	0.969	-1.37
glaucoma	G1	42	missingness	0.25	-0.046	0.092	-0.148
glaucoma	G1	42	missingness	0.50	-0.028	0.046	-0.061
glaucoma	G1	42	missingness	0.75	-0.0100	0.019	-0.067
glaucoma	G1	42	shuffle_train	train	-0.011	-0.050	-0.058
glaucoma	G1	73	capacity	random_d6	-0.080	0.969	-1.06
glaucoma	G1	73	missingness	0.25	0.011	0.090	0.017
glaucoma	G1	73	missingness	0.50	0.023	0.0067	0.0097
glaucoma	G1	73	missingness	0.75	-0.0049	-0.022	0.0007
glaucoma	G1	73	shuffle_train	train	-0.079	0.029	-0.138
glaucoma	G2	13	capacity	random_d6	-0.706	0.688	-0.162
glaucoma	G2	13	missingness	0.25	-0.012	-0.212	-0.108
glaucoma	G2	13	missingness	0.50	-0.0065	-0.300	-0.079
glaucoma	G2	13	missingness	0.75	0.0073	-0.433	-0.041
glaucoma	G2	13	shuffle_train	train	-0.0034	-0.488	-0.103
glaucoma	G2	42	capacity	random_d6	-0.717	0.688	-0.060
glaucoma	G2	42	missingness	0.25	-0.0012	-0.056	-0.098
glaucoma	G2	42	missingness	0.50	0.0063	-0.127	0.0024
glaucoma	G2	42	missingness	0.75	0.0051	-0.210	-0.0018
glaucoma	G2	42	shuffle_train	train	0.0080	-0.491	-0.020
glaucoma	G2	73	capacity	random_d6	-0.648	0.706	-0.049
glaucoma	G2	73	missingness	0.25	0.0074	-0.093	-0.043
glaucoma	G2	73	missingness	0.50	-0.0001	-0.320	-0.032
glaucoma	G2	73	missingness	0.75	0.0011	-0.298	-0.046
glaucoma	G2	73	shuffle_train	train	0.011	-0.307	-0.091
glaucoma	G3	13	capacity	random_d6	-0.706	0.688	-0.162
glaucoma	G3	13	missingness	0.25	-0.012	-0.212	-0.108
glaucoma	G3	13	missingness	0.50	-0.0065	-0.300	-0.079
glaucoma	G3	13	missingness	0.75	0.0073	-0.433	-0.041
glaucoma	G3	13	shuffle_train	train	-0.0034	-0.488	-0.103
glaucoma	G3	42	capacity	random_d6	-0.733	0.740	-0.067
glaucoma	G3	42	missingness	0.25	-0.026	-0.123	-0.152
glaucoma	G3	42	missingness	0.50	-0.011	-0.119	0.0078
glaucoma	G3	42	missingness	0.75	-0.0062	-0.207	-0.047
glaucoma	G3	42	shuffle_train	train	0.013	-0.360	-0.141
glaucoma	G3	73	capacity	random_d6	-0.648	0.706	-0.049
glaucoma	G3	73	missingness	0.25	0.0074	-0.093	-0.043
glaucoma	G3	73	missingness	0.50	-0.0001	-0.320	-0.032
glaucoma	G3	73	missingness	0.75	0.0011	-0.298	-0.046
glaucoma	G3	73	shuffle_train	train	0.011	-0.307	-0.091
avmnist	A0	13	capacity	random_d12544	-0.012	–	0.130
avmnist	A0	13	missingness	0.25	0.014	-0.125	0.208
avmnist	A0	13	missingness	0.50	0.0032	-0.049	0.157
avmnist	A0	13	missingness	0.75	-0.0068	-0.025	0.091
avmnist	A0	13	shuffle_train	train	-0.014	-0.012	0.056
avmnist	A0	42	capacity	random_d12544	-0.016	–	0.120
avmnist	A0	42	missingness	0.25	0.015	-0.172	0.208
avmnist	A0	42	missingness	0.50	0.0039	-0.066	0.154
avmnist	A0	42	missingness	0.75	-0.0066	-0.014	0.087
avmnist	A0	42	shuffle_train	train	-0.015	-0.010	0.053
avmnist	A0	73	capacity	random_d12544	-0.016	–	0.120
avmnist	A0	73	missingness	0.25	0.013	-0.133	0.209
avmnist	A0	73	missingness	0.50	0.0038	-0.055	0.160
avmnist	A0	73	missingness	0.75	-0.0055	-0.020	0.092
avmnist	A0	73	shuffle_train	train	-0.017	-0.011	0.050
avmnist	A0	101	capacity	random_d12544	-0.013	–	0.124
avmnist	A0	101	missingness	0.25	0.013	-0.130	0.206
avmnist	A0	101	missingness	0.50	0.0025	-0.049	0.153
avmnist	A0	101	missingness	0.75	-0.0073	-0.033	0.085
avmnist	A0	101	shuffle_train	train	-0.017	-0.018	0.050
avmnist	A0	137	capacity	random_d12544	-0.015	–	0.121
avmnist	A0	137	missingness	0.25	0.014	-0.128	0.209
avmnist	A0	137	missingness	0.50	0.0036	-0.050	0.155
avmnist	A0	137	missingness	0.75	-0.0059	-0.021	0.096

Dataset	Level	Seed	Control	Setting	IMG	RS	RMU
avmnist	A0	137	shuffle_train	train	-0.016	-0.014	0.054
avmnist	A3	13	capacity	random_d12544	-0.074	0.019	-0.079
avmnist	A3	13	missingness	0.25	0.0024	-0.0008	-0.0043
avmnist	A3	13	missingness	0.50	0.0003	-0.0012	-0.0046
avmnist	A3	13	missingness	0.75	-0.0047	-0.0016	-0.0077
avmnist	A3	13	shuffle_train	train	-0.0090	-0.0062	-0.013
avmnist	A3	42	capacity	random_d12544	-0.073	0.019	-0.078
avmnist	A3	42	missingness	0.25	0.0008	-0.0004	-0.0057
avmnist	A3	42	missingness	0.50	-0.0018	-0.0003	-0.0065
avmnist	A3	42	missingness	0.75	-0.0051	-0.0005	-0.0081
avmnist	A3	42	shuffle_train	train	-0.010	-0.0060	-0.015
avmnist	A3	73	capacity	random_d12544	-0.078	0.019	-0.083
avmnist	A3	73	missingness	0.25	0.0012	0.0002	-0.0059
avmnist	A3	73	missingness	0.50	-0.0019	-0.0008	-0.0071
avmnist	A3	73	missingness	0.75	-0.0046	-0.0010	-0.0076
avmnist	A3	73	shuffle_train	train	-0.012	-0.0063	-0.017
avmnist	A3	101	capacity	random_d12544	-0.074	0.019	-0.080
avmnist	A3	101	missingness	0.25	0.0026	0.0006	-0.0040
avmnist	A3	101	missingness	0.50	-0.0001	-0.0000	-0.0050
avmnist	A3	101	missingness	0.75	-0.0042	-0.0004	-0.0073
avmnist	A3	101	shuffle_train	train	-0.0098	-0.0063	-0.014
avmnist	A3	137	capacity	random_d12544	-0.074	0.019	-0.079
avmnist	A3	137	missingness	0.25	0.0027	-0.0008	-0.0039
avmnist	A3	137	missingness	0.50	0.0003	-0.0013	-0.0047
avmnist	A3	137	missingness	0.75	-0.0023	-0.0018	-0.0051
avmnist	A3	137	shuffle_train	train	-0.011	-0.0061	-0.015
mosi	T0	13	capacity	random_d5	0.055	0.759	-0.0093
mosi	T0	13	missingness	0.25	0.039	-0.010	0.074
mosi	T0	13	missingness	0.50	0.033	-0.0059	0.075
mosi	T0	13	missingness	0.75	0.013	0.0010	0.033
mosi	T0	13	shuffle_train	train	0.0016	-0.060	0.0015
mosi	T0	42	capacity	random_d5	0.045	0.759	-0.0019
mosi	T0	42	missingness	0.25	0.046	-0.0078	0.073
mosi	T0	42	missingness	0.50	0.011	0.0002	0.025
mosi	T0	42	missingness	0.75	-0.024	-0.0056	-0.033
mosi	T0	42	shuffle_train	train	0.0080	-0.054	0.0076
mosi	T0	73	capacity	random_d5	0.056	0.756	0.022
mosi	T0	73	missingness	0.25	0.038	-0.0045	0.049
mosi	T0	73	missingness	0.50	0.016	-0.0099	0.023
mosi	T0	73	missingness	0.75	-0.0002	-0.0026	-0.0005
mosi	T0	73	shuffle_train	train	0.0057	-0.042	0.0059
mosi	T0	101	capacity	random_d5	0.031	0.756	-0.0031
mosi	T0	101	missingness	0.25	0.059	-0.0017	0.086
mosi	T0	101	missingness	0.50	0.038	-0.0074	0.052
mosi	T0	101	missingness	0.75	0.025	-0.012	0.034
mosi	T0	101	shuffle_train	train	0.022	-0.049	0.032
mosi	T0	137	capacity	random_d5	0.022	0.759	0.0069
mosi	T0	137	missingness	0.25	0.047	-0.0059	0.076
mosi	T0	137	missingness	0.50	0.033	-0.0056	0.043
mosi	T0	137	missingness	0.75	0.025	-0.0051	0.044
mosi	T0	137	shuffle_train	train	-0.058	-0.049	-0.104
mosi	T1	13	capacity	random_d5	0.021	0.977	-0.0031
mosi	T1	13	missingness	0.25	0.013	-0.035	0.029
mosi	T1	13	missingness	0.50	0.017	-0.041	0.032
mosi	T1	13	missingness	0.75	0.0025	-0.020	0.012
mosi	T1	13	shuffle_train	train	0.0024	-0.091	0.0028
mosi	T1	42	capacity	random_d5	0.020	0.977	-0.0027
mosi	T1	42	missingness	0.25	0.019	-0.023	0.036
mosi	T1	42	missingness	0.50	0.011	-0.023	0.022
mosi	T1	42	missingness	0.75	-0.013	-0.050	-0.020
mosi	T1	42	shuffle_train	train	-0.0005	-0.092	-0.0013
mosi	T1	73	capacity	random_d5	0.016	0.977	-0.055
mosi	T1	73	missingness	0.25	0.018	-0.030	0.018
mosi	T1	73	missingness	0.50	0.0066	-0.052	0.0052
mosi	T1	73	missingness	0.75	-0.0026	-0.052	-0.018
mosi	T1	73	shuffle_train	train	-0.0021	-0.070	-0.014
mosi	T1	101	capacity	random_d5	0.016	0.976	-0.071
mosi	T1	101	missingness	0.25	0.036	-0.026	0.052
mosi	T1	101	missingness	0.50	0.023	-0.039	0.033
mosi	T1	101	missingness	0.75	0.018	-0.049	0.029
mosi	T1	101	shuffle_train	train	0.012	-0.078	0.019
mosi	T1	137	capacity	random_d5	0.017	0.977	-0.048
mosi	T1	137	missingness	0.25	0.021	-0.017	0.027
mosi	T1	137	missingness	0.50	0.015	-0.027	0.015
mosi	T1	137	missingness	0.75	-0.0001	-0.029	-0.0063
mosi	T1	137	shuffle_train	train	-0.030	-0.087	-0.065
mosi	T2	13	capacity	random_d5	-0.0035	0.923	0.0005
mosi	T2	13	missingness	0.25	0.0003	-0.0070	0.0008
mosi	T2	13	missingness	0.50	-0.0004	-0.0065	-0.0009

Dataset	Level	Seed	Control	Setting	IMG	RS	RMU
mosi	T2	13	missingness	0.75	0.0005	0.0021	0.0026
mosi	T2	13	shuffle_train	train	-0.0007	-0.055	-0.0016
mosi	T2	42	capacity	random_d5	-0.0042	0.923	0.0002
mosi	T2	42	missingness	0.25	0.0015	-0.0063	0.0020
mosi	T2	42	missingness	0.50	0.0006	-0.0042	0.0019
mosi	T2	42	missingness	0.75	-0.0001	-0.014	0.0008
mosi	T2	42	shuffle_train	train	0.0010	-0.045	0.0005
mosi	T2	73	capacity	random_d5	-0.0039	0.924	-0.0021
mosi	T2	73	missingness	0.25	0.0012	-0.0058	0.0019
mosi	T2	73	missingness	0.50	0.0007	-0.0095	0.0015
mosi	T2	73	missingness	0.75	0.0004	-0.0036	0.0010
mosi	T2	73	shuffle_train	train	0.0016	-0.054	0.0022
mosi	T2	101	capacity	random_d5	-0.0053	0.922	0.0022
mosi	T2	101	missingness	0.25	0.0003	-0.011	-0.0009
mosi	T2	101	missingness	0.50	0.0006	-0.016	0.0003
mosi	T2	101	missingness	0.75	0.0004	-0.014	-0.0005
mosi	T2	101	shuffle_train	train	0.0004	-0.048	0.0001
mosi	T2	137	capacity	random_d5	-0.0030	0.924	-0.0004
mosi	T2	137	missingness	0.25	0.0032	-0.0037	0.0057
mosi	T2	137	missingness	0.50	0.0023	-0.0084	0.0035
mosi	T2	137	missingness	0.75	0.0015	-0.0088	0.0024
mosi	T2	137	shuffle_train	train	-0.0013	-0.057	-0.0000
mosi	T3	13	capacity	random_d5	-0.0026	0.958	-0.0000
mosi	T3	13	missingness	0.25	0.0004	-0.012	-0.0005
mosi	T3	13	missingness	0.50	0.0008	-0.013	-0.0008
mosi	T3	13	missingness	0.75	-0.0000	0.0032	-0.0008
mosi	T3	13	shuffle_train	train	-0.0021	-0.058	-0.0036
mosi	T3	42	capacity	random_d5	-0.0018	0.958	-0.0003
mosi	T3	42	missingness	0.25	0.0002	0.0019	-0.0013
mosi	T3	42	missingness	0.50	0.0004	-0.014	-0.0002
mosi	T3	42	missingness	0.75	0.0005	-0.021	0.0017
mosi	T3	42	shuffle_train	train	0.0006	-0.057	0.0009
mosi	T3	73	capacity	random_d5	-0.0033	0.958	-0.0000
mosi	T3	73	missingness	0.25	0.0008	-0.0027	0.0009
mosi	T3	73	missingness	0.50	0.0003	-0.0087	0.0002
mosi	T3	73	missingness	0.75	-0.0001	-0.011	-0.0005
mosi	T3	73	shuffle_train	train	-0.0012	-0.053	-0.0027
mosi	T3	101	capacity	random_d5	-0.0033	0.957	0.0003
mosi	T3	101	missingness	0.25	-0.0006	-0.013	-0.0023
mosi	T3	101	missingness	0.50	-0.0013	-0.015	-0.0023
mosi	T3	101	missingness	0.75	-0.0005	-0.016	-0.0014
mosi	T3	101	shuffle_train	train	-0.0013	-0.055	-0.0025
mosi	T3	137	capacity	random_d5	-0.0016	0.959	0.0002
mosi	T3	137	missingness	0.25	0.0018	-0.0046	0.0010
mosi	T3	137	missingness	0.50	0.0011	-0.0078	0.0003
mosi	T3	137	missingness	0.75	-0.0010	-0.0080	-0.0032
mosi	T3	137	shuffle_train	train	-0.0015	-0.068	-0.0018

F Statistical details

All correlations are Spearman ρ with seed-cluster bootstrap (pooled analyses cluster by dataset×seed), $B = 10{,}000$, fixed bootstrap seeds, percentile 95% intervals. The pre-specified gate requires a positive point estimate and an interval excluding zero. The secondary mixed model standardises all variables within dataset and fits

$$\text{IMG} \sim \text{RMU} + \text{RS} + \text{level} + \text{dataset intercepts}$$

by OLS with cluster-bootstrap intervals: RMU coefficient 0.467 (cluster CI [0.231,0.751]), RS coefficient 0.107, and level coefficient -0.569. With only 12 representation conditions it is descriptive; we report effect sizes and intervals rather than p-values, and no main-text conclusion relies on it.

G Residual baseline head-to-head and probe robustness

Matched comparison. Every statistic is computed on the same frozen cells with identical splits, train-only standardisation, validation-selected nuisance heads g, h on the same ridge grids, the same $K = 5$ out-of-fold folds and fold seeds, and the same held-out residuals. Each method contributes one pre-specified scalar to the identical gate. The scalars are: *partial correlation*, mean absolute sample correlation over all residual pairs; *residual* R^2, a

descriptive in-sample R^2 of an $\alpha = 1$ ridge direction measured within the held-out residual sample; *PLS-1*, held-out risk reduction of the single NIPALS latent component fitted on training residuals; *CCA*, first regularised canonical correlation with ridge $\lambda = 1/n$ on both sides and symmetric whitening (Hotelling, 1936); and *HSIC* (centred kernel alignment) with linear and Gaussian median-bandwidth kernels (Gretton et al., 2005). Gaussian HSIC is not computed for AV-MNIST (10,000 held-out points). The RMU re-computation here reproduces the committed per-cell CSV; the re-derived gate matches the main-text gate to the reported precision.

Table 5: Head-to-head gates $\rho(\text{statistic}, \text{IMG})$ [95% bootstrap CI] under the fully matched protocol. RMU has the strongest pooled rank association; PLS-1 is close; the unsupervised dependence statistics are markedly weaker.

Statistic	Glaucoma	AV-MNIST	CMU-MOSI	Pooled
RMU (cross-fitted ridge probe)	0.845 [0.333, 1.000]	0.767 [0.599, 0.876]	0.974 [0.919, 0.992]	0.651 [0.483, 0.825]
Partial correlation (mean abs.)	0.592 [0.333, 1.000]	0.717 [0.596, 0.879]	0.482 [0.280, 0.767]	0.419 [0.317, 0.567]
Residual R^2 (in-sample, descriptive)	0.451 [0.333, 0.482]	0.903 [0.834, 0.971]	0.385 [0.265, 0.684]	0.270 [0.182, 0.414]
PLS-1 probe (held-out)	0.310 [0.200, 0.401]	0.679 [0.420, 0.861]	0.947 [0.900, 0.981]	0.620 [0.377, 0.847]
Regularized CCA (first)	0.458 [0.333, 0.538]	0.887 [0.793, 0.972]	0.362 [0.242, 0.630]	0.273 [0.161, 0.494]
HSIC, linear kernel	0.683 [0.333, 1.000]	0.583 [0.226, 0.866]	0.371 [0.242, 0.653]	0.388 [0.234, 0.544]
HSIC, Gaussian kernel	0.528 [0.400, 1.000]	–	0.841 [0.768, 0.940]	0.483 [0.355, 0.824]

Probe robustness. The held-out residual probe is replaced by ElasticNet, PLS with one and two components, and a seeded small MLP (one 32-unit hidden layer, Adam, fixed epochs); all are fit on training residuals only, with the gate, bootstrap, and clustering unchanged. Table 6 shows the result. Linear low-capacity probes are consistent with ridge (PLS-2 numerically strongest, PLS-1 close). The ElasticNet failure on AV-MNIST uses a five-pass SGD fit because dense coordinate descent is intractable at this scale; it yields large negative risk reductions on several AV-MNIST cells and near-zero values on glaucoma regression cells (strong L_1 attenuation)—an estimator-convergence/attenuation failure, reported as such rather than read as an absence of signal. The small-MLP gate is weak, so the claim is bounded to reasonable linear low-capacity probe families.

Table 6: Probe-family robustness of the $\rho(\text{RMU}, \text{IMG})$ gate; same residuals, splits, bootstrap, and clustering.

Probe	Glaucoma	AV-MNIST	CMU-MOSI	Pooled
Ridge	0.845 [0.333, 1.000]	0.767 [0.599, 0.876]	0.974 [0.919, 0.992]	0.651 [0.484, 0.829]
ElasticNet†	0.444 [-0.333, 1.000]	0.429 [-0.058, 0.729]	-0.755 [-0.919, -0.617]	0.033 [-0.355, 0.355]
PLS-1	0.310 [0.200, 0.401]	0.679 [0.420, 0.861]	0.947 [0.900, 0.981]	0.620 [0.377, 0.847]
PLS-2	0.676 [0.333, 1.000]	0.741 [0.595, 0.876]	0.917 [0.695, 0.994]	0.713 [0.546, 0.882]
Small MLP	0.528 [0.333, 1.000]	-0.298 [-0.809, 0.132]	0.267 [-0.200, 0.669]	0.291 [-0.024, 0.540]

Sensitivity to weakest levels. Excluding all weakest-level cells (G0/A0/T0) attenuates every method: RMU to 0.369 $[0.023, 0.711]$, with Gaussian HSIC the only other interval excluding zero (0.313 $[0.038, 0.674]$); partial correlation, residual R^2, CCA, and linear HSIC intervals then span zero. Excluding G0 only leaves RMU at 0.582 $[0.378, 0.777]$. Weak representations therefore contribute disproportionately to *every* diagnostic's gate, not only RMU's.

Methodological audits. Three dated audits accompany this revision: the nearest-neighbour novelty audit, the cached-artifact inventory and reuse decisions, and the data-access protocol audit, which reconstructs every test-split touch. Within the frozen ladders the test split is read once per cell for evaluation, and test metrics logged in encoder cache files are post-hoc with no feedback. Three pre-freeze development pivots (the discarded AV-MNIST encoder family, the A2 epoch schedule, and the MOSI GRU conditioning choice) are disclosed with timestamps and registry evidence; no frozen decision used test results.

H Matched backbone sweep matrix

The full matrix from the independently executed matched backbone sweep (test split; seed-mean ± seed SD, in dB) covers all seven vision representations, four text encoders, and raw/redacted conditions. It is generated from the committed `results_multiseed` aggregate CSVs by `research_2027/iclr/analysis/make_mega_panel.py`; the exact Shapley values and patient-group bootstrap statistics follow that protocol's own contract.

Vision	Text encoder	Condition	ϕ_{OCT} (mean±SD)	ϕ_{TEXT} (mean±SD)
V1_convnext_ft	T1_bioclinicalbert	raw	+0.124 ± 0.003	-0.000 ± 0.002
V1_convnext_ft	T1_bioclinicalbert	red	+0.120 ± 0.005	+0.024 ± 0.007
V1_convnext_ft	T2_biomedbert	raw	+0.118 ± 0.005	+0.018 ± 0.002
V1_convnext_ft	T2_biomedbert	red	+0.125 ± 0.005	+0.023 ± 0.005
V1_convnext_ft	T3_deberta_v3	raw	+0.115 ± 0.007	+0.017 ± 0.003
V1_convnext_ft	T3_deberta_v3	red	+0.115 ± 0.005	+0.023 ± 0.005
V1_convnext_ft	T4_modernbert	raw	+0.112 ± 0.004	+0.053 ± 0.002
V1_convnext_ft	T4_modernbert	red	+0.114 ± 0.005	+0.043 ± 0.001
V2_retfound_frozen	T1_bioclinicalbert	raw	+0.202 ± 0.006	-0.016 ± 0.004
V2_retfound_frozen	T1_bioclinicalbert	red	+0.206 ± 0.004	+0.005 ± 0.004
V2_retfound_frozen	T2_biomedbert	raw	+0.196 ± 0.006	+0.018 ± 0.005
V2_retfound_frozen	T2_biomedbert	red	+0.205 ± 0.009	+0.015 ± 0.004
V2_retfound_frozen	T3_deberta_v3	raw	+0.206 ± 0.004	+0.005 ± 0.006
V2_retfound_frozen	T3_deberta_v3	red	+0.200 ± 0.005	+0.011 ± 0.001
V2_retfound_frozen	T4_modernbert	raw	+0.194 ± 0.007	+0.015 ± 0.003
V2_retfound_frozen	T4_modernbert	red	+0.197 ± 0.006	+0.018 ± 0.006
V3_retfound_pft2	T1_bioclinicalbert	raw	+0.007 ± 0.006	-0.011 ± 0.003
V3_retfound_pft2	T1_bioclinicalbert	red	+0.005 ± 0.004	-0.002 ± 0.002
V3_retfound_pft2	T2_biomedbert	raw	+0.005 ± 0.001	+0.005 ± 0.003
V3_retfound_pft2	T2_biomedbert	red	+0.008 ± 0.004	+0.000 ± 0.003
V3_retfound_pft2	T3_deberta_v3	raw	+0.005 ± 0.005	+0.017 ± 0.002
V3_retfound_pft2	T3_deberta_v3	red	+0.008 ± 0.003	+0.020 ± 0.002
V3_retfound_pft2	T4_modernbert	raw	+0.007 ± 0.004	+0.013 ± 0.004
V3_retfound_pft2	T4_modernbert	red	+0.006 ± 0.003	+0.008 ± 0.002
V4_dinov2_frozen	T1_bioclinicalbert	raw	+0.358 ± 0.004	-0.009 ± 0.007
V4_dinov2_frozen	T1_bioclinicalbert	red	+0.360 ± 0.007	+0.009 ± 0.006
V4_dinov2_frozen	T2_biomedbert	raw	+0.351 ± 0.005	+0.019 ± 0.006
V4_dinov2_frozen	T2_biomedbert	red	+0.358 ± 0.005	+0.015 ± 0.005
V4_dinov2_frozen	T3_deberta_v3	raw	+0.350 ± 0.004	+0.020 ± 0.005
V4_dinov2_frozen	T3_deberta_v3	red	+0.346 ± 0.003	+0.028 ± 0.006
V4_dinov2_frozen	T4_modernbert	raw	+0.340 ± 0.006	+0.019 ± 0.006
V4_dinov2_frozen	T4_modernbert	red	+0.342 ± 0.004	+0.023 ± 0.005
V5_dinov2_pft2	T1_bioclinicalbert	raw	+0.015 ± 0.005	-0.007 ± 0.002
V5_dinov2_pft2	T1_bioclinicalbert	red	+0.015 ± 0.004	+0.004 ± 0.004
V5_dinov2_pft2	T2_biomedbert	raw	+0.016 ± 0.005	+0.007 ± 0.001
V5_dinov2_pft2	T2_biomedbert	red	+0.013 ± 0.005	+0.010 ± 0.003
V5_dinov2_pft2	T3_deberta_v3	raw	+0.014 ± 0.002	+0.011 ± 0.003
V5_dinov2_pft2	T3_deberta_v3	red	+0.015 ± 0.004	+0.000 ± 0.002
V5_dinov2_pft2	T4_modernbert	raw	+0.014 ± 0.004	+0.019 ± 0.005
V5_dinov2_pft2	T4_modernbert	red	+0.015 ± 0.005	+0.021 ± 0.007
V6_retclip_tower	T1_bioclinicalbert	raw	+0.180 ± 0.008	-0.005 ± 0.002
V6_retclip_tower	T1_bioclinicalbert	red	+0.177 ± 0.007	+0.015 ± 0.001
V6_retclip_tower	T2_biomedbert	raw	+0.174 ± 0.002	+0.001 ± 0.003
V6_retclip_tower	T2_biomedbert	red	+0.171 ± 0.002	+0.012 ± 0.005
V6_retclip_tower	T3_deberta_v3	raw	+0.178 ± 0.005	+0.018 ± 0.002
V6_retclip_tower	T3_deberta_v3	red	+0.178 ± 0.006	+0.014 ± 0.001
V6_retclip_tower	T4_modernbert	raw	+0.171 ± 0.005	+0.020 ± 0.004
V6_retclip_tower	T4_modernbert	red	+0.165 ± 0.004	+0.013 ± 0.002
V7_eyeclip_tower	T1_bioclinicalbert	raw	+0.677 ± 0.002	-0.020 ± 0.006
V7_eyeclip_tower	T1_bioclinicalbert	red	+0.665 ± 0.002	+0.027 ± 0.005
V7_eyeclip_tower	T2_biomedbert	raw	+0.656 ± 0.002	-0.005 ± 0.002
V7_eyeclip_tower	T2_biomedbert	red	+0.659 ± 0.002	+0.000 ± 0.005
V7_eyeclip_tower	T3_deberta_v3	raw	+0.705 ± 0.001	-0.059 ± 0.003
V7_eyeclip_tower	T3_deberta_v3	red	+0.684 ± 0.002	-0.037 ± 0.001

Vision	Text encoder	Condition	ϕ_{OCT} (mean±SD)	ϕ_{TEXT} (mean±SD)
V7_eyeclip_tower	T4_modernbert	raw	+0.648 ± 0.003	+0.019 ± 0.005
V7_eyeclip_tower	T4_modernbert	red	+0.646 ± 0.002	+0.026 ± 0.002

I Synthetic identification experiments

Three seeded DGPs are committed as unit tests, with 6,000 samples each, ridge at fixed penalty and cross-fitted residuals: (i) *no signal*: under $Y \perp A \mid Z$ with $A = f(Z) + \varepsilon_A$, RS is high while IMG and RMU are statistically zero; (ii) *positive residual*: with $A = f(Z) + \varepsilon_A$ and $Y = 2Z + \beta\varepsilon_A + \varepsilon_Y$ ($\beta = 1$), both IMG and RMU are positive; (iii) *scaling*: RMU is strictly larger at $\beta = 1.5$ than at $\beta = 0$, matching the population expression in Appendix A. The regression suite passes in the committed environment; the analogous seeded classification checks used during development (null RMU 0.004 under conditional independence; 0.021 with injected residual label information) are documented in the project status records.

J Reproducibility manifest

- Software: Python 3.10.20; torch 2.3.1+cu121; scikit-learn 1.7.2; scipy 1.15.3; pandas 2.3.3; numpy 2.2.6; matplotlib 3.10.9. Hardware: NVIDIA RTX 3090 (24 GB); float64 ridge normal equations on CPU.
- Frozen configurations: `configs/avmnist_ladder.yaml`, `configs/mosi_ladder.yaml`, and `configs/glaucoma_ladder.yaml`. Seeds: glaucoma $\{13, 42, 73\}$; AV-MNIST and MOSI $\{13, 42, 73, 101, 137\}$; encoder seed offsets 1000/2000/3000; bootstrap seeds 20260910/20260911, B=10,000.
- Glaucoma split: 454/83/90 patient groups = 1,125/208/222 records, with zero group overlap asserted by test. An internal split manifest (`audits/glaucoma_split_manifest.json`) records group-set identifiers for private audit only; the hash values are not reproduced in this paper.
- CMU-MOSI `mosi_data.pkl` MD5 `63f092700a4565d5208bcc6afa187552`, canonical 1,284/229/686; AV-MNIST upstream MultiBench pack with /255 scaling.
- Registry: 12 glaucoma + 60 AV + 60 MOSI ladder runs and 210 control runs, each recording run id, split manifest id, seed, config id, and metrics in `research_2027/results/experiment_registry.csv`.
- Logged single-GPU wall-clock (registry timestamps, last complete row per cell; duplicate append-only rows deduplicated; scheduling not normalized): glaucoma ladder 18.4 min; CMU-MOSI 46.5 min; AV-MNIST 30.2 h; controls 101.0 h; approximately 132 h over the 342 unique runs. No new training was performed for this revision; all statistics derive from the frozen artifacts above.

K Submission statements

Generative-AI use. Generative AI tools were used solely to assist with language polishing, grammatical correction, and improvements to the clarity and readability of the manuscript. All research ideas, methodological design, implementations, experiments, data collection and analysis, interpretation of results, figures and tables, and scientific conclusions were independently conducted, verified, and approved by the authors. The authors take full responsibility for the accuracy, integrity, and content of the manuscript.

Ethics. The clinical data (fundus images, OCT measures, visual fields) are a retrospective, de-identified cohort; patient-group splits prevent patient leakage, and no raw clinical data, identifiers, or images are redistributed. The cohort has no external or cross-device validation; nothing here claims clinical readiness or diagnostic deployment. The public benchmarks (AV-MNIST, CMU-MOSI) contain no personal health information and are used under

their research licences. The work proposes an analysis diagnostic, not autonomous clinical decision-making. The applicable institutional ethics/governance statement is provided by the authors in the submission form.

Reproducibility. All splits are fixed and documented (a glaucoma patient-group split manifest with private audit hash values not printed in this paper; MultiBench canonical 55k/5k/10k and 1,284/229/686 splits). Seeds derive encoder initialisation, permutations, missingness masks, cross-fitting folds, and Gaussian capacity inputs. Hyperparameter grids and fixed epoch schedules live in committed YAML configs; penalties are selected on validation only; within the frozen protocol each cell reads the fixed held-out split once (the clinical split has prior project use and is not claimed as untouched external validation). Every cell has a registry row with run id, split manifest id, config id, and metrics. Aggregate tables, statistics, and figures are generated by committed analysis code from aggregate CSVs. Public-dataset features are not redistributed under research-use licensing; loaders, source locations, versions, and verified split sizes are committed so results can be re-fetched. The anonymous supplementary ZIP contains the full recompute code, configs, safe aggregate data, synthetic identification tests, and generated figures. The package reproduces all reported statistical analyses, tables, and figures from safe aggregate artifacts, and public-benchmark models can be retrained end-to-end from the configs; raw clinical data and private clinical checkpoints/embeddings are not redistributed, so the clinical representation-extraction stage is not independently reproducible from the public package.