A MODALITY WITH NO NEW INFORMATION CAN IMPROVE MULTIMODAL FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

A fusion model that beats its strongest unimodal component is often credited with using information from the added modality. Adding a branch, however, also changes parameterization and optimization. We propose a *rotation-placebo audit* that separates the two: beside the real modality, retrain the same fusion architecture with an invertible orthogonal rotation of the raw base input, which adds zero Shannon information, and with parameter-matched copy and Gaussian arms. A gain that the placebo reproduces cannot be attributed to new information. In the knowledge-tracing deployment that motivated this work, the audit overturns the original attribution: the placebo raises area under the curve by an amount comparable to the gain that had justified collecting student feedback, including records from minors. It also raises it in the public Computer Science Education Data Mining (CSEDM) challenge, though there its log-loss gain ($+0.005$ bits) is below our 0.01-bit floor. The audit is also specific. It stays near zero where a real modality helps substantially: on frozen Contrastive Language–Image Pretraining (CLIP) features, in a registered audit of the official Adaptive Gradient Modulation implementation, and in a second public audit that reproduces a $+10.81$-point real-modality gain while copy and rotation gain only $+1.96$ and $+0.65$ points. Where the placebo fires, registered interventions control it. It reduces held-out cross-entropy in all 12 registered raw-input audio–visual MNIST (AV-MNIST) configurations (60 paired runs). Saturating the task removes it, and rank-256 principal-component truncation and whitening make all six settings fail our prespecified positive-placebo criterion while preserving a positive real-modality gain (largest residual $+0.022$ bits, positive in 4/5 seeds). A converged probe adds a one-sided certificate that bounds the cost of ignoring a candidate, with no violations among 42 eligible cases. A base-versus-fusion comparison alone cannot attribute a gain to new information; the audit makes that attribution testable.

1 INTRODUCTION

Multimodal systems are commonly justified by an ablation: train a base model, add a modality, retrain, and call the difference a multimodal gain. This evidence can motivate new sensors, storage, production pipelines, and collection of sensitive data. It is therefore often interpreted as showing that the added modality supplied useful information unavailable to the base. The ablation does not isolate that claim.

Adding an input changes both the available information and the learner's parameterization and optimization path. We separate them with an orthogonal rotation of the base representation. This invertible input preserves all base information while presenting it through a second path. We call it the *rotation placebo* and its measured improvement the *rotation gain* g_{rot}. A positive rotation gain cannot come from new information in the candidate.

Here, "registered" means that the directional prediction and success criterion were fixed before observing the outcome. In the knowledge-tracing (KT) deployment that motivated this work, the placebo gain is comparable to the feedback gain that had justified collecting records from minors (Figure 1B). The evidence does not show that the feedback is worthless; it shows that the original ablation did not establish why the system improved.

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

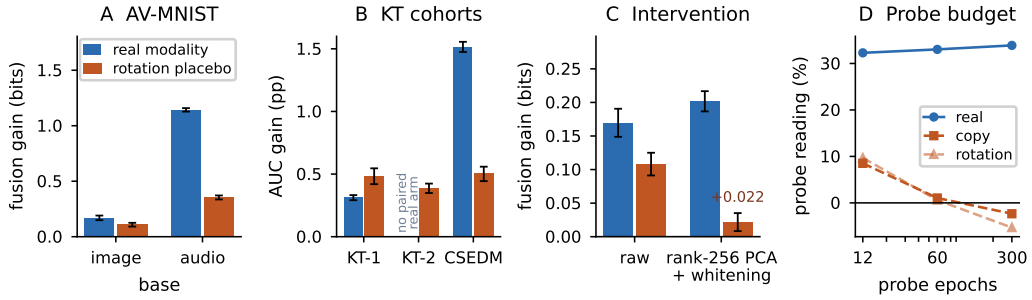


Figure 1: **The rotation-placebo audit fires, and registered interventions switch it off.** "Rotation" is an orthogonal re-encoding of the raw base followed by per-block scaling; "real" is the paired modality. Bars are paired means ± standard error over seeds (A, C) or folds (B). **A:** both candidates improve AV-MNIST fusion from either base. **B:** the placebo raises AUC in three KT cohorts; KT-1 shows the shared-base five-fold follow-up, and KT-2 has no paired real arm. **C:** after rank-256 PCA truncation and whitening (early-stopping cell), the real gain remains; the residual rotation gain, $+0.022$ bits in 4/5 seeds, fails the criterion. **D:** longer probe training drives copy and rotation readings to or below zero (pooled over seven cells) while the real reading stays near 33%.

The audit is not a detector that always fires. On frozen CLIP features, in the official AGM implementation, and in an independent public late-fusion audit, rotation and copy stay near zero or small while the real modality helps substantially. Where the placebo does fire, registered interventions control it: it improves all 12 registered raw-input AV-MNIST configurations (60 paired runs), saturation switches it off, restoring a noisy coordinate tail switches it on, and combined rank-256 principal-component analysis (PCA) truncation and whitening make all six settings fail the positive-placebo criterion while preserving positive real-modality gains. Registered misses rule out simple accounts based only on optimizer choice, sample size, or label noise. Section 6 reports the paired estimates and uncertainty.

We therefore report the rotation placebo beside the unimodal baseline. The real-modality gain asks whether the augmented system improves; the placebo-adjusted gain asks whether it improves beyond an information-equivalent reparameterization. A separate converged probe can issue a *one-sided ignore certificate*, which bounds the cost of ignoring the candidate. The certificate is valid only after the probe reads zero on copy and rotation controls. A large probe reading does not promise that fusion will exploit the candidate.

We make three contributions. First, we propose the rotation-placebo audit, with copy and Gaussian arms reported beside the unimodal baseline, and show that it overturns the attribution of a deployed fusion gain. Second, we show that the audit is specific: it stays null where real modalities help substantially (frozen CLIP, official AGM, independent late fusion) and fires where registered interventions leave optimization headroom. Third, we derive a linear optimization-relief account, use its empirical failures to delimit it, and pair the audit with a one-sided ignore certificate that has no violations among the 42 eligible cases.

2 Related Work: Attribution and Optimization

Multimodal fusion and foundation representations. Multimodal benchmarks can reward fusion even when complex architectures do not beat strong unimodal baselines or use little cross-modal interaction (Hessel & Lee, 2020; Rheude et al., 2026). This concern spans cross-modal attention (Tsai et al., 2019; Nagrani et al., 2021; Jia et al., 2024) and pretrained vision–language representations (Kim et al., 2021; Radford et al., 2021; Singh et al., 2022). Prior comparisons ask which architecture wins; we ask what a win establishes about information in the added input.

Is a multimodal gain actually multimodal? The closest precedent, empirical multimodally-additive function projection (EMAP), removes cross-modal interactions after training and finds that much of the performance survives (Hessel & Lee, 2020). Our placebo instead fixes incremental

information to zero before retraining. EMAP tests whether a learned prediction needs interaction; we test whether a training gain needs new information. Both show that comparison with a unimodal baseline is insufficient.

Multimodal optimization and modality imbalance. Joint training can underuse a modality because streams learn at different rates (Wang et al., 2020). Gradient modulation, Adaptive Gradient Modulation (AGM), Prototypical Modal Rebalance, and MMPareto address this imbalance (Peng et al., 2022; Li et al., 2023; Fan et al., 2023; Wei & Hu, 2024). They show that training dynamics are not fixed by the information set. We study the converse attribution error: dynamics can improve even when the candidate contains no new information.

Controls, shortcuts, and benchmark variance. Sanity controls and shortcut studies show that plausible measurements need not track their intended mechanism (Adebayo et al., 2018; Geirhos et al., 2020). Our control preserves the base signal exactly. To separate the effect from seed and split variance (Bouthillier et al., 2021; Picard, 2021), we use paired seeds, fixed splits, registered interventions, and retained negative settings.

Reparameterization, normalization, and noise. Equivalent functions can have very different parameter-space geometry (Dinh et al., 2017), overparameterization can accelerate optimization without enlarging the function class (Arora et al., 2018), and wide networks can develop redundant groups carrying the same signal with independent noise (Doimo et al., 2022). These results provide context for our exact linear redundant-path result. Section 4 measures how far that result extends empirically. Normalization changes the smoothness of optimization (Santurkar et al., 2018) and whitening changes feature covariance, preventing representation collapse (Ermolov et al., 2021); our train-only rank-256 PCA truncation and whitening therefore cannot identify either component alone. Input noise can itself regularize (Bishop, 1995), but matched Gaussian blocks do not reproduce our effect: the distinction that matters is a signal-dependent deterministic re-encoding, not width or arbitrary noise. Our implementation audit shows that per-block standardization makes the effective rotation anisotropic. The follow-up isolates a reproducible effect of that preprocessing, although it does not fully explain the rotation–copy gap.

Information accessibility. Predictive $\mathcal{V}$-information formalizes what a restricted predictor family can actually use, so an invertible computation can change accessibility (Xu et al., 2020), and conditional probing measures what is usable beyond a baseline (Hewitt et al., 2021). This does not invalidate our claim: the rotation leaves both the Shannon quantity and, for families closed under orthogonal input maps, the accessible function class unchanged (Section 3, Appendix A), and Section 7 gives the empirical counterpart.

Modality selection and acquisition. Modality selection and partial information decomposition quantify redundancy, uniqueness, and synergy (Cheng et al., 2022; He et al., 2024; Liang et al., 2023a; Xiu et al., 2026). Recent controlled analysis also separates task-composition and fusion bottlenecks in multimodal reasoning (Wang et al., 2026). These studies characterize predictions or reasoning failures; our control asks whether a retrained fusion gain requires incremental information. Active acquisition decides which measurements to purchase at inference time (Valancius et al., 2024; Rheude et al., 2025), MultiBench systematizes evaluation across modalities and fusion methods (Liang et al., 2023b), and transferability estimators score frozen representations before fine-tuning (Nguyen et al., 2020; You et al., 2021; Huang et al., 2022). These characterize what modalities contain, how they are acquired, or how they transfer. Our result concerns the ruler: realized fusion gain can be positive when incremental information is fixed at zero.

3 What a rotation placebo controls for

The controls separate candidate information from improvement created by an augmented learner.

3.1 AN INFORMATION-EQUIVALENT INPUT

Let $X \in \mathbb{R}^d$ be the base representation and let Q be an orthogonal matrix fixed before training. The placebo candidate is

$$X_{\mathrm{rot}} = XQ, \qquad Q^\top Q = I. \tag{1}$$

Because XQ is a deterministic function of X, $I(X_{\mathrm{rot}}; Y \mid X) = 0$ for every distribution; invertibility also ensures that the rotation loses no information from X. Predictive $\mathcal{V}$-information (Xu et al., 2020) can differ under a restricted family, but for any family closed under invertible linear input maps, such as a dense first layer that absorbs the map into its weights, the accessible class is also unchanged (Appendix A). We train matched base and fusion architectures with $[X, X_{\mathrm{rot}}]$ or $[X, M]$ for a real modality M. **Implementation-critical preprocessing.** The implementation fits a separate standard scaler, fitted on the training split, to each input block. Thus Equation 1 describes the information-preserving raw input, while the network receives a centered, coordinate-rescaled version of each block. This preprocessing preserves the zero-incremental-information control but need not preserve orthogonality. Dense candidate branches use separate input projections, so XQ and M may have different widths; a structured branch must also decode its control input (Appendix C).

For held-out loss L, define

$$g_{\mathrm{real}} = L(f_X) - L(f_{X,M}), \tag{2}$$

$$g_{\mathrm{rot}} = L(f_X) - L(f_{X,XQ}). \tag{3}$$

The real-modality gain compares two trained systems but does not by itself show that M supplied new information. We call $g_{\mathrm{real}} - g_{\mathrm{rot}}$ the *placebo-adjusted gain*: improvement beyond an information-equivalent candidate.

3.2 COPY, ROTATION, AND GAUSSIAN ARMS ISOLATE THE REDUNDANT PATH

We use three controls. COPY appends X itself; ROT appends XQ; and GAUSS appends independent Gaussian columns matched to the width of X, not necessarily M. Copy and rotation contain no incremental information; Gaussian noise matches width and parameters without providing a redundant signal path. We pair all arms by split, seed, architecture, optimizer, and budget.

Our registered representation intervention combines rank-256 PCA truncation and whitening; it is not whitening alone. We fit the transform on training data only and apply it unchanged to held-out data, and repeat all arms. It jointly removes the discarded coordinate tail and rescales retained components. It can therefore test whether the combined intervention removes g_{rot}, but cannot attribute that change to truncation or whitening separately. Throughout, we say that an intervention *removes* the rotation gain when the resulting setting no longer meets the positive-placebo criterion of Section 5; we do not claim that it is exactly zero. The residual gain and its standard error appear in Table 1.

3.3 A PROBE IS VALID ONLY AFTER ITS COPY-NULL CONVERGES

A second instrument estimates the conditional log-loss reduction available from a candidate. Small linear and multilayer-perceptron (MLP) probes receive out-of-fold predictions of the deployed base, available base features, and optionally the candidate. For KT, the base features include outcome, skill-difficulty, and history summaries; public-data probes use train-fitted PCA coordinates (Appendix C). Their held-out cross-entropy gap is

$$\widehat{I}_p(M; Y \mid X) = \mathrm{CE}_p(X) - \mathrm{CE}_p(X, M). \tag{4}$$

A probe is valid only after it passes a copy-null gate: $\widehat{I}_p(X; Y \mid X) \approx 0$ and $\widehat{I}_p(XQ; Y \mid X) \approx 0$. A nonzero null reading means that the augmented parameterization is improving the probe itself. The remedy is to converge or retune the probe, not to subtract the observed floor. The recorded certificate audit applies one eligibility check: the base probe must beat the marginal predictor. Out-of-fold base predictions are a design requirement, and copy-null convergence is a separate diagnostic prerequisite for using the converged probe. We do not retrospectively count these requirements as additional filters on the 42 eligible certificates.

The interpretation is one-sided: under explicit approximation slack, a small reading bounds the cost of ignoring a candidate, whereas a large reading does not guarantee that a chosen fusion architecture will exploit it.

4 A LINEAR ACCOUNT OF PLACEBO GAINS AND ITS LIMITS

The placebo does not change the Bayes predictor, but it changes an optimized finite model. The linear case can be computed exactly.

Proposition 1 (Redundant-path relief)**.** *Consider a linear predictor on X with penalty $\lambda\|w\|^2$. Append a re-encoding XA and parameterize the same prediction as $X\beta_1 + XA\beta_2$. The smallest penalty that realizes base direction w is*

$$\inf_{\beta_1 + A\beta_2 = w} \lambda(\|\beta_1\|^2 + \|\beta_2\|^2) = \lambda w^\top (I + AA^\top)^{-1} w. \tag{5}$$

For a copy or square orthogonal rotation in this linear parameterization, the effective penalty is exactly halved.

Standardization makes the rotation anisotropic. That orthogonal special case does not describe the standardized rotation used by the deep experiments. If D_b and D_r contain the training-set coordinate scales of the base and rotated blocks, then after centering the second branch is related to the first by $A = D_b Q D_r^{-1}$. For raw-image AV-MNIST, a post hoc code audit gives condition number 1.22×10^4, so the eigenvalues of $I + AA^\top$ span 1.00 to 19.40. Copy still has $A = I$ and exact halving, but rotation supplies anisotropic relief. This makes per-block scaling a candidate source of the rotation–copy asymmetry, not an explanation by itself. Across isotropic, covariance, and fitted-direction weightings, the relief factor ranges from 0.485 to 0.627 versus 0.500 for copy, so the advantage is small or reverses sign, far short of the observed deep-model gap. An unregistered paired follow-up then replaces the pipeline map by an exactly orthogonal map after base standardization. The original pipeline rotation has a larger gain in all 10 seed–budget pairs (exact one-sided $p = .031$ within each five-seed budget), but a frozen secondary copy-gap criterion fails under early stopping and one orthogonal run is unstable. Per-block scaling therefore affects the magnitude and stability of the effect without providing a complete causal account (Appendix B).

Where the linear account stops. A redundant path can reduce the penalty for the same predictor when the original regularization is suboptimal. The inequality $(I + AA^\top)^{-1} \preceq I$ establishes this direction, but does not imply that every finite training run improves. In nonlinear networks, optimization speed may play a similar role. Two measurements delimit how far Proposition 1 carries to our deep models. First, the registered parameter-matched copy arm matches branch count and parameters, yet rotation exceeds copy in 28 of 30 AV-MNIST comparisons (Section 6). Where the placebo stays near zero, copy and rotation are also nearly equal. Their AV-MNIST gap is consistent with the anisotropic standardized map above, although these experiments do not establish that conditioning alone causes the nonlinear effect. Second, Equation 5 predicts exactly zero relief at $\lambda = 0$, yet the registered weight-decay sweep measures 0.108 bits there. The gain is positive in 5/5 seeds and the largest in the sweep to within 0.0002 bits; predicted and measured dependence on λ run in opposite directions. The proposition therefore explains the direction of relief under explicit regularization but not the deep-model magnitude or its dependence on λ; Appendix A gives the proof.

Where tested, the placebo therefore depends on exploitable optimization slack. Our interventions associate that slack with a strong learnable signal, a noisy coordinate tail, and an unsaturated task. These are empirical conditions, not an if-and-only-if theorem.

Proposition 2 (One-sided ignore certificate)**.** *Let f_X be the deployed base and $f_{X,M}$ any fusion predictor. Let $\epsilon = \mathrm{CE}(f_X) - H(Y \mid X)$ be base fitting slack and let δ bound the probe's asymmetric regret between X and (X, M). Then*

$$\mathrm{CE}(f_X) - \mathrm{CE}(f_{X,M}) \le I(M; Y \mid X) + \epsilon \le \widehat{I}_p(M; Y \mid X) + \delta + \epsilon. \tag{6}$$

A small right-hand side certifies that ignoring M has bounded regret. Exchanging X and M gives the corresponding bound for switching to M. The converse is false: a large reading does not certify that fusion will help. In the audit, $\hat{\epsilon}$ is a prespecified 2% allowance and $\hat{\delta}$ is the gap between

linear and MLP base probes; the latter is a heuristic proxy, not a proven bound. We apply the certificate eligibility check and probe requirements in Section 3.3. Section 7 audits coverage against independently trained fusion models.

5 A REGISTERED SEQUENCE OF DISCRIMINATING EXPERIMENTS

The registered sequence tests where the placebo appears and why.

Domains. AV-MNIST from MultiBench pairs 28×28 digit images with audio spectrograms (Liang et al., 2023b); either raw modality can serve as the base. Cross-domain negative settings use the Carnegie Mellon University Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) and Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) datasets (Zadeh et al., 2016; 2018), Hateful Memes (Kiela et al., 2020), and the Canadian Institute for Cybersecurity Intrusion Detection System 2017 (CICIDS2017) dataset (Sharafaldin et al., 2018). Three knowledge-tracing (KT) cohorts use recurrent experts over student histories, code, and feedback; two are private and one is the public CSEDM 2019 Data Challenge (Price et al., 2019). Appendix C describes all features and splits.

Architectures and evaluation. Public-data models are a gated fusion network and a two-layer, four-head cross-attention model. Each uses fixed 5- and 20-epoch budgets plus early stopping with a 200-epoch ceiling, with five paired seeds per configuration. The headline KT experiment uses a recurrent mixture of experts and five held-out folds. We report held-out cross-entropy and the task metric without selecting seeds.

Registered sequence. "Registered" means the prediction and success criterion were written into a timestamped internal experiment ledger before the run. Table 3 in Appendix B records the frozen sequence, its preserved misses, and the unregistered shared-base follow-up. Appendix B gives every prediction and outcome; the supplement also includes the earlier criterion-development runs.

Decision criteria. A setting meets the *positive-placebo criterion* if its paired mean cross-entropy gain exceeds 0.01 bits and it meets the registered replication threshold (every seed for the raw-input headline, 5/5 for the official audit); task-metric-only audits retain their directional rule. The floor was fixed before E7; re-scoring at 0.005 and 0.02 bits changes no AV-MNIST, PCA, or CLIP conclusion (Table 4).

6 WHERE THE PLACEBO FIRES AND WHERE IT STAYS NULL

6.1 THE PLACEBO FIRES IN RAW-INPUT AND DEPLOYMENT SETTINGS

An orthogonal rotation of the raw base contains exactly zero incremental information, yet it improves the fused model in every registered image-base AV-MNIST run (Table 1). It does so in all six registered architecture–budget cells and all five seeds per cell. It reproduces 44–108% of the image-base real-modality improvement across these cells. An unregistered five-matrix follow-up remains positive in all 50 runs. Against a parameter-matched copy, rotation is larger in 15/15 gated-fusion and 13/15 cross-attention runs; mean copy gain is 31% and 48.4% of rotation gain, respectively. Matched Gaussian features usually hurt. Branch multiplicity and generic noise therefore do not explain the larger rotation gain within these AV-MNIST settings; part of the excess reflects per-block standardization (Section 4).

The effect is not specific to the image side. With the raw audio spectrogram as the base and the image as the paired real modality, using a direction fixed before the run, the rotation placebo is positive in all six architecture–budget combinations and in every seed. Across cells its mean gain ranges from $+0.22$ to $+1.49$ bits, compared with $+1.14$ to $+4.51$ bits for the paired real image. The corresponding cellwise ratios range from 15% to 33%. Accuracy agrees, so the gain is not a log-loss calibration artifact: rotation raises test accuracy in all 60 runs, by $+0.57$ to $+1.28$ points from the image base and $+5.98$ to $+9.78$ from the audio base. Rotation also exceeds the matched copy in all 30 audio-base runs. This comparison was measured in the registered control arms but was not a separate registered success criterion.

The audit overturns the attribution in the motivating recurrent knowledge-tracing (KT) deployment. In the registered three-fold comparison (E19), real and rotation gains are $+0.273$ and $+0.515$ area under the receiver operating characteristic curve (AUC) percentage points, each positive in 3/3 folds; the arms retrain their bases, so the difference is descriptive (corrected $p = .151$). An unregistered five-fold follow-up sharing the trained base gives $+0.312$ and $+0.483$, each positive in 5/5 folds (corrected $p = .105$; Appendix B) (Nadeau & Bengio, 2003).

A second private cohort yields $+0.387$ AUC pp. In the public CSEDM 2019 Data Challenge data (Price et al., 2019), the same trained base is shared across arms: the real-code gain is $+1.515$ AUC percentage points and the rotation gain is $+0.501$ AUC percentage points ($+0.0049$ bits, below our 0.01 floor), both positive in 3/3 folds. CSEDM has no feedback observations, so this run validates the real-code and missing-input paths; three folds limit exact inference.

Frozen foundation representations and transformer fusion. On frozen CLIP (Radford et al., 2021) features from Hateful Memes, a trainable cross-attention transformer gains 0.0993 bits from real text (5/5 seeds), while rotation and copy remain near zero (0.0063 and 0.0093 bits) and Gaussian features hurt. Real text remains positive after the PCA intervention. Thus the controls distinguish the real modality in this compact pretrained representation.

Setting	Candidates	Mean gain (metric)	Direction
AV-MNIST, image base	real / copy / rot.	$+.05$–$.37/+.01$–$.12/+.04$–$.40$ bits	rot. 30/30; rot.>copy 28/30
AV-MNIST, audio base	real / rot.	$+1.14$–$+4.51/+.22$–$+1.49$ bits	rot. 30/30
KT cohort 1 (E19)	real / rot.	$+0.273/+0.515$ AUC pp	both 3/3 folds
KT cohort 2	rot.	$+0.387$ AUC pp	3/3 folds
Public CSEDM[‡]	real / rot.	$+1.515/+0.501$ AUC pp	both 3/3 folds
Frozen CLIP	real / copy / rot.	$+.0993/.0093/.0063$ bits	5/5, 4/5, 2/5
Official joint training[§]	real / copy / rot.	$+5.70/-.05/+.04$ accuracy pp	5/5, 3/5, 3/5
Official AGM[§]	real / copy / rot.	$+6.73/+.07/+.04$ accuracy pp	5/5, 3/5, 3/5
MLA late fusion[†]	real / copy / rot.	$+10.81/+1.96/+.65$ accuracy pp	5/5, 4/5, 3/5
Rank-256 PCA AV-MNIST	real / rot.	$+.10$–$.21/-.121$–$+.022$ bits	all 6 settings fail criterion

Table 1: Where the zero-incremental-information placebo fires, and where it does not. "rot." is rotation, "pp" is percentage points of the named task metric, and MLA is Multimodal Learning with Alternating Unimodal Adaptation; cross-metric magnitudes are not comparable. Gains use matched splits and protocols. [§]Official AV-MNIST joint-training and AGM implementations use the same split and protocol. [†]Unregistered; MLA direction counts use accuracy, with cross-entropy directions reported in Appendix B; [‡]the same trained base is shared across arms. Appendix B reports registration status and all misses. Cross-row ratios are denominator-driven, not a severity ladder. The rank-256 PCA row spans six cells; its largest is the cell plotted in Figure 1C, $+0.022$ bits (standard error 0.013, 4/5 seeds): failure of the criterion does not mean an exactly zero gain.

6.2 Where tested, the placebo tracks a learned base, headroom, and a noisy tail

Pooled MOSI/MOSEI, raw-sequence variants, and CICIDS2017 (Sharafaldin et al., 2018) do not establish the effect. Registered interventions suggest three empirical conditions; a post hoc spectrum audit does not support a monotone cross-dataset spectral account (Appendix B).

A strong learnable base signal. The placebo is absent for our weak bag-of-words, pooled, and raw-sequence bases. The AV-MNIST audio base, by contrast, reaches 0.62–0.67 accuracy against 0.10 chance even where its loss exceeds marginal entropy through overconfidence. The relevant condition is a learned signal to re-encode, not low loss; our registered text-transfer prediction missed.

Remaining task headroom. Restricting AV-MNIST to digits 0 versus 1 reduces base cross-entropy and rotation gain to about 0.002 bits each: saturation switches the effect off. Adding label noise does not; that registered prediction missed.

A noisy coordinate tail. At width 784, rank-k AV-MNIST bases show no placebo for $k \leq 128$; restoring the full tail at $k = 784$ yields $+0.104$ bits under early stopping and $+0.372$ at 20 epochs, in every seed (Figure 2B). After rank-256 PCA truncation and whitening, all six settings fail the

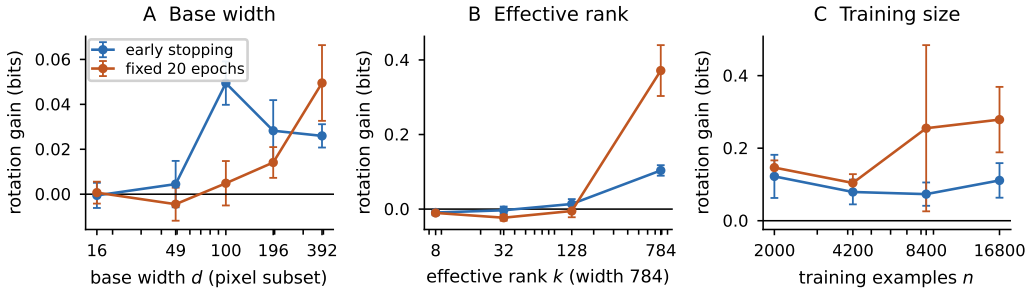


Figure 2: **The placebo needs more than width or data.** Each panel varies one quantity while holding its stated protocol fixed. All panels use gated fusion, AdamW, an experiment-specific rotation Q, and five-seed standard errors. Their horizontal axes are logarithmic. Vertical ranges differ: A spans about 0.06 bits against 0.4–0.5 in B and C. **A:** width raises the gain but does not switch it on. The fixed budget rises overall (Spearman $+0.90$, exact two-sided permutation $p = .083$, $n = 5$) while early stopping peaks at $d = 100$ and falls back ($+0.60$, $p = .350$), so the registered width experiment has a mixed outcome. Its $d = 784$ cell is omitted as not protocol-identical and kept in Appendix B. **B:** effective rank does switch it. At width 784, rank-k reconstructions show no placebo for $k \leq 128$; the effect appears only once the full noisy tail returns, in 5/5 seeds at both budgets. **C:** growing the training set from 2,000 to 16,800 examples does not shrink the effect; the wide interval at $n = 8{,}400$ is one high-variance cell.

positive-placebo criterion while the real-modality gain remains positive, though it declines 43.6% in the largest-effect cell. The combined intervention cannot isolate its components.

Stochastic gradient descent (SGD) enlarges rather than removes the placebo (registered miss), so the effect is not specific to AdamW. It also persists with more data and across the tested weight-decay grid. Four small penalties barely change base loss and the base optimum is at the grid edge; larger penalties remain untested (Appendix B). These three conditions describe tested dense inputs, not an if-and-only-if law. The tail intervention was AV-MNIST-only; branch-decodability is a separate architectural prerequisite.

6.3 Two external audits preserve large real-modality gains

The registered AV-MNIST audit (Li et al., 2023) preserves the official model and protocol. Real audio gains $+5.70$ accuracy percentage points under joint training and $+6.73$ under AGM (5/5 seeds), whereas rotation and rerun copy stay near zero (Table 1; Appendix B). Rotation destroys pixel locality before the audio residual network (ResNet) (He et al., 2016), so this miss also marks an architectural boundary; accuracy alone cannot rule out a log-loss gain.

An unregistered audit of the independently published Multimodal Learning with Alternating Unimodal Adaptation (MLA) (Zhang et al., 2024) late-fusion configuration on the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) (Cao et al., 2014) provides a second public test. Across five paired seeds, real audio raises accuracy by 10.81 ± 2.19 pp (5/5), whereas independently trained copy and rotation branches gain 1.96 ± 2.73 and 0.65 ± 0.99 pp. Rotation is smaller than copy in 4/5 pairs. Cross-entropy gives the same ordering: $+0.364$, $+0.101$, and $+0.067$ bits. Thus the audit preserves the large published real-modality effect while identifying the smaller null gains as generic branch ensembling rather than a rotation-specific reproduction (Appendix B.2).

7 Using the audit and interpreting its scope

Run the attribution arms before interpreting the gain. Fit the paired base, real-modality, copy, rotation, and Gaussian arms in that order, using the same split, initialization, architecture, optimizer, and budget. First verify that the branch can decode its information-equivalent control; then report the placebo-adjusted gain. A claim that the candidate can be safely ignored additionally requires a converged probe and its copy-null check.

Claim	Paired unit	Reported inference	Multiplicity
AV-MNIST placebo	seed	direction, mean, standard error	no joint p claim
KT shared base	held-out fold	sign-flip; correction	tests separate
Frozen CLIP	seed	paired bootstrap interval	descriptive, $n = 5$
Official AGM	seed	direction; copy sign-flip	threshold 5/5
MLA late fusion	paired seed	direction, mean, standard deviation	exploratory, $n = 5$
Operating curves	swept-cell mean	exact Spearman test	no family-wise claim
Certificates	candidate	0 violations among 42 eligible	deterministic bound

Table 2: Inference is matched to the experimental unit. We report claim families separately and do not claim family-wise error control across experiments.

The probe can certify when its null controls converge. Pooled over seven cells in three domains, 12-epoch probes read $+8.51\%$ on copy and $+9.69\%$ on rotation. At 300 epochs the pooled null readings are non-positive (-2.34%, -5.32%), but copy stays positive in 4/7 cells and rotation in 3/7, with three cell readings above our 2% allowance, while the pooled real reading remains about 33% of base uncertainty. The copy-null gate is therefore a live prerequisite, not a formality; the registered prediction that fusion would behave likewise missed.

Of 161 candidate cells, 106 receive certificates and 42 pass the eligibility gate recorded at protocol v1.0, before the certificate audit was run; none of those 42 is violated. The three violations lie outside that gate, in cells whose base probe is weaker than the marginal predictor. A small valid reading bounds ignore loss, but a large one does not guarantee fusion.

Table 2 states the statistical unit and multiplicity of each claim; Appendices B and C give full outcomes and implementation details.

External outcomes delimit rather than estimate prevalence. This protocol tests attribution, not how often placebo gains occur. CREMA-D suggests a decodability boundary: its locality-preserving cyclic permutation fires through a convolutional ResNet-18, whereas the registered official-AGM audit's locality-destroying dense transform stays null before the same encoder class. In the MLA audit, copy exceeds rotation and both remain far below the real-modality gain. These post hoc contrasts are not causal. The PCA intervention conflates truncation and whitening; we claim no family-wise error control (Appendix B).

8 Conclusion

A base-versus-fusion comparison cannot, on its own, attribute a gain to new information; an added input can improve a trained system without adding any. The rotation-placebo audit makes the attribution testable. In the motivating deployment it overturned the original attribution; on frozen CLIP, official AGM, and MLA it stayed null or small while a real modality helped substantially; and in raw-input AV-MNIST registered interventions switched it off and on. The audit therefore distinguishes attribution rather than presuming failure.

It is an audit, not a correction factor: the placebo tests whether an information-equivalent input reproduces a gain, copy and Gaussian arms separate re-encoding from branch count and noise, and a converged copy-null gate licenses the ignore certificate.

This distinction matters whenever an ablation supports a downstream decision. A positive fusion gain establishes that the augmented system performed better. It does not, on its own, establish that collecting, retaining, or deploying the new modality caused that improvement through new information. Reporting the information-equivalent controls makes the stronger claim testable and exposes cases in which architecture or optimization supplies an alternative explanation. The audit requires no new labels and preserves negative outcomes as useful scope evidence. Its cost is three additional arms and one converged probe per attribution claim, so it scales with model size.

Before attributing a gain to another modality, test whether the same pipeline improves with an input known to add no information.

AI USE STATEMENT

We used generative AI tools to scaffold code, draft analysis scripts that aggregate the experiment ledger, and generate figures from that ledger. The tools also assisted with literature discovery, language editing, and drafts of the related-work and discussion sections and figure captions.

They also assisted with mathematical and experimental drafts. The authors checked the propositions and proofs in Appendix A, chose the final hypotheses and success criteria, and fixed each registered prediction in the timestamped experiment ledger before observing its corresponding outcome. Registered misses are preserved in Table 3 exactly as scored. No generative AI tool was used to generate synthetic datasets; every dataset is a released benchmark or a de-identified deployment log.

The authors reviewed all AI-assisted work and independently checked the generated code and prose. Every reported number was verified against retained raw records or frozen aggregate artifacts; Appendix C discloses the external-audit raw records that were not retained. We take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

The deployment experiment uses de-identified interaction logs, including records from minors, to evaluate models rather than participants. No identity or free-text record appears in the paper or supplement. Private cohorts are not released. The paper reports negative results and a data-minimization implication: when a sensitive modality has no measurable incremental value, not collecting it may be preferable. Public benchmark experiments use their standard released features.

REPRODUCIBILITY STATEMENT

Sections 3–5 define the controls, hypotheses, and evaluation protocol. The appendix records all registered predictions, including misses, and gives architectures, budgets, and data splits. The submitted supplement contains the experiment ledger, retained raw and aggregate records, analysis code, and scripts used to generate the figures and tables. Most original seed-0–3 raw files of the E24 audit were lost with their execution host; Appendix C lists what is retained and what was rerun.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, volume 31, 2018.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *ICML*, pp. 244–253, 2018.

Christopher M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Runxiang Cheng, Gargi Balasubramaniam, Yifei He, Yao-Hung Hubert Tsai, and Han Zhao. Greedy modality selection via approximate submodular maximization. In *UAI*, 2022.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *ICML*, pp. 1019–1028, 2017.

Diego Doimo, Aldo Glielmo, Sebastian Goldt, and Alessandro Laio. Redundant representations help generalization in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, pp. 3015–3024, 2021.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *CVPR*, pp. 20029–20038, 2023.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. Efficient modality selection in multimodal learning. *Journal of Machine Learning Research*, 25, 2024.

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*, pp. 861–877, 2020. doi: 10.18653/v1/2020.emnlp-main.62.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. Conditional probing: Measuring usable information beyond a baseline. In *EMNLP*, pp. 1626–1639, 2021. doi: 10.18653/v1/2021.emnlp-main.122.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *ICML*, 2022.

Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. GeminiFusion: Efficient pixel-wise multimodal fusion for vision transformer. In *ICML*, pp. 21753–21767, 2024.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*, pp. 5583–5594, 2021.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pp. 22214–22224, 2023.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. In *NeurIPS*, 2023a.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Ziyin Wu, Yun Cheng, Jason Wu, Leslie Y. Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. MultiBench: Multiscale benchmarks for multimodal representation learning. *Journal of Machine Learning Research*, 24(370):1–48, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52: 239–281, 2003. doi: 10.1023/A:1024068626366.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, volume 34, 2021.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cédric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *ICML*, 2020.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pp. 8238–8247, 2022.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

Thomas W. Price et al. CSEDM 2019 data challenge. Second Educational Data Mining in Computer Science Education Workshop, at LAK 2019, 2019. Dataset documentation.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Tillmann Rheude, Roland Eils, and Benjamin Wild. Cohort-based active modality acquisition. *arXiv preprint arXiv:2505.16791*, 2025.

Tillmann Rheude, Roland Eils, and Benjamin Wild. Fusion or confusion? multimodal complexity is not all you need. *arXiv preprint arXiv:2512.22991*, 2026. Version 3.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *NeurIPS*, volume 31, 2018.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *International Conference on Information Systems Security and Privacy*, pp. 108–116, 2018. doi: 10.5220/0006639801080116.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, pp. 15638–15650, 2022.

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 *ACL*, pp. 6558–6569, 2019. doi: 10.18653/v1/P19-1656.

Michael Valancius et al. Acquisition conditioned oracle for nongreedy active feature acquisition. In *ICML*, 2024.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pp. 12695–12705, 2020.

Yucheng Wang, Yifan Hou, Aydin Javadov, Mubashara Akhtar, and Mrinmaya Sachan. Compose and fuse: Revisiting the foundational bottlenecks in multimodal reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=oIvIsK5AwB.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*, pp. 52559–52572, 2024.

Lixin Xiu, Xufang Luo, and Hideki Nakayama. A comprehensive information-decomposition analysis of large vision-language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6WsBGk4Iag.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *ICLR*, 2020.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *ICML*, 2021.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACL*, 2018.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27456–27466, 2024.

A Theory and proofs

A.1 Proof of Proposition 1

For a fixed realized direction w, write $\beta_1 = w - A\beta_2$. The minimum is therefore

$$\min_z \|w - Az\|^2 + \|z\|^2.$$

The normal equation gives $z = (I + A^\top A)^{-1} A^\top w$. Substitution and the push-through identity $A(I + A^\top A)^{-1} = (I + AA^\top)^{-1} A$ yield

$$\min_z \|w - Az\|^2 + \|z\|^2 = w^\top (I + AA^\top)^{-1} w.$$

For $A = I$ or square orthogonal $A = Q$, $I + AA^\top = 2I$, so the effective quadratic penalty is halved. Independent Gaussian columns are not of the form XA and the argument does not apply.

A.2 Proof of Proposition 2

No predictor using (X, M) can have expected cross-entropy below $H(Y \mid X, M)$. Hence

$$\begin{aligned} \mathrm{CE}(f_X) - \mathrm{CE}(f_{X,M}) &\le H(Y \mid X) + \epsilon - H(Y \mid X, M) \\ &= I(M; Y \mid X) + \epsilon. \end{aligned}$$

If δ upper-bounds the difference between the two probe approximation regrets, then $I(M; Y \mid X) \le \widehat{I}_p(M; Y \mid X) + \delta$, which gives Equation 6. The inequality is one-sided: neither step lower-bounds the gain of a particular fusion learner.

A.3 Why invertibility is sufficient

For any fixed invertible linear map A, XA is a deterministic function of X, so $I(XA; Y \mid X) = 0$. Conversely $X = (XA)A^{-1}$, so X and XA have the same sigma-algebra and the same Bayes risk. Any observed difference between their finite trained systems must therefore arise from estimation, optimization, or regularization rather than additional population information. The predictive $\mathcal{V}$-information of Xu et al. (2020) is not an exception here: if the predictor family is closed under invertible linear input maps, then composition with A^{-1} is a bijection of the family, so attainable risk and $\mathcal{V}$-information are unchanged. A dense first layer gives this closure by absorbing A^{-1} into its weights; the spatial ResNet of E24 does not, which is why that null also tests encoder compatibility.

B Registered experiments and complete outcomes

Table 3 records the completed discriminating sequence. A "miss" means the registered directional prediction failed; misses were not used to redefine success. In the table, Rot is the rotation arm, Copy the unrotated duplicate, and Gauss the matched independent-Gaussian arm.

Registration notes. E11's 3/3 comprises two additional KT cohorts and the AV-MNIST audio-base placement, so all 30 audio-base runs are registered. E9 is a registered placebo-only run with a separately trained base; E19 adds a registered real arm, and E22b is the unregistered shared-base five-fold follow-up. E20 was dropped without an observed result under its recorded one-day environment-risk clause. E8 was registered as "PCA-whiten" without a retained rank in the prose ledger; pre-execution code fixed $k = 256$ and whitening. E13 was scored on all six registered

ID	Question	Prediction	Outcome
E7	Four-domain null controls	positive $\geq 3/4$; vanishes converged	both miss; AV-MNIST rot. $>$ copy 15/15
E8	Rank-256 truncation + whitening	rot. near zero in all 6; real survives	hit; all 6 fail criterion
E9	KT placebo, per-arm base	placebo $+.05$–$.4$ pp; probe $<.01$ bits	above band; probe partial (1/3)
E10	Raw MOSI/MOSEI sequences	positive transfer	miss, 1/12
E11	Additional KT cohorts	positive transfer	hit, 3/3
E12	CICIDS intrusion features	placebo $+$; intervention removes; real $+$	placebo miss; later tests not run
E13	Within-data width dose	AV-MNIST monotone; CICIDS on	AV-MNIST hit; subset mixed; CICIDS miss
E14	AdamW specificity	SGD smaller	miss
E15a	Label-noise headroom	restores placebo	miss
E15b	Saturated binary task	removes placebo	hit
E16	Rank at fixed width	grows with rank or stays flat	grows only at full rank, all seeds
E17	Sample size	effect shrinks with n	miss
E18	Bag-of-words text	positive transfer	miss; base invalid
E19	Fold-aligned KT, per-arm base	real in band, not above placebo	both hit; real $+.273$, rot. $+.515$
E20	Published MultiBench code	run unless setup exceeds one day	dropped unrun under risk clause
E21	Frozen CLIP + fusion transformer	real positive, placebo null	hit; real 5/5, rotation 2/5
E22a	CSEDM shared-base audit	real positive	hit; real larger 3/3
E22b	KT shared-base, five folds	unregistered follow-up	both positive 5/5
E23	AdamW weight-decay sweep	tuned base closes placebo	miss; 6/6 settings positive
E24	Official AGM audit	rotation positive 5/5	miss: real 5/5, rot 3/5; copy rerun 3/5

Table 3: Directional outcomes with registration status preserved. E1–E6 are criterion-development and convergence experiments in the ledger. Notes below explain rows whose registration or execution status needs qualification.

widths ($\rho = .94/.77$). Figure 2A excludes the protocol-distinct $d = 784$ endpoint and reports the five-width subset ($\rho = .90/.60$), which is why its caption calls the outcome mixed. E16's registered alternatives concerned intrinsic dimension versus headroom. The "noisy tail" interpretation followed the full-rank switch and is not labeled as the registered prediction.

Official AGM audit (E24). E24 preserves the official split, architecture, optimizer, 100-epoch budget, and selection rule. Joint training yields $+5.70$ accuracy percentage points for real audio and $+0.04$ for rotation; AGM yields $+6.73$ and $+0.04$. Real fusion is positive in 5/5 seeds, whereas rotation is positive in 3/5 and misses the registered 5/5 criterion.

Recovered E24 copy arm. The copy arm was rerun with frozen E24 code after its original raw records were lost with the execution host. Its mean gains are -0.054 points under joint training (95% paired-bootstrap interval $[-0.225, 0.117]$) and $+0.072$ under AGM ($[-0.109, 0.245]$), with 3/5 positive seeds in each setting. The supplement retains the five paired values and exact sign-flip tests.

Shared-base knowledge tracing (E22). E22b shares the trained base between arms. Real and rotation gains are $+0.312$ and $+0.483$ AUC points, both positive in 5/5 folds; their difference is not significant (exact $p = .063$; corrected $p = .105$). For public CSEDM, the real-modality gain exceeds the placebo in all three folds; the exact test is resolution-limited at $p = .125$.

Copy comparison (E7). E7 matches architecture and parameters across copy and rotation. Rotation exceeds copy in 15/15 gated-fusion and 13/15 cross-attention comparisons; mean copy gains are 31% and 48.4% of rotation gains. This rejects branch multiplicity alone, not every form of ensemble regularization. With the audio base, rotation exceeds copy in all 30 measured runs; this was not a separately registered success criterion.

Intervention qualifications. In the largest-effect PCA cell, the real-modality gain falls from 0.366 to 0.206 bits but remains positive; this 43.6% decrease is not generalized to other cells. The E23 weight-decay grid provides little effective contrast and ends at the measured base optimum, so it does not establish a dose response. Figure 2A omits the protocol-distinct E13 $d = 784$ cell ($0.069/0.129$ bits); the E16 $k = 784$ endpoint is $0.104/0.372$ bits. Both remain in the supplement.

Encoder and spectrum audits. COPY and ROT share reshaping, upsampling, and the official audio ResNet; only ROT applies the fixed re-encoding. A post hoc spectrum audit does not support a monotone cross-dataset account, so we make no spectral causal claim.

Standardized-rotation follow-up (E26, unregistered). After identifying the per-block standard scaler, we specified a paired follow-up before execution. It compares COPY, the original independently standardized rotation (ROT-PIPE), and an exactly orthogonal rotation applied after base standardization (ROT-ORTHO), using gated fusion, five seeds, and the fixed-20 and early-stopping budgets. ROT-PIPE exceeds ROT-ORTHO in 5/5 seeds under each budget (exact one-sided $p = .031$). Mean gains for COPY/ROT-ORTHO/ROT-PIPE are $+.168/-.025/+.370$ bits at fixed 20 and $+.004/-.166/+.066$ under early stopping. The frozen primary direction succeeds, but the requirement that ROT-ORTHO be closer to COPY in both budgets fails under early stopping because one orthogonal run is unstable. We therefore score the composite prediction as a miss and treat scaling as a contributor, not a complete explanation. The negative ROT-ORTHO means are driven by seed 1 (-0.790 and -0.833 bits): the medians are $+0.102$ and $+0.008$ bits. At fixed 20, excluding that disclosed seed gives ROT-ORTHO $+0.166$ and COPY $+0.163$ bits; ROT-PIPE still exceeds ROT-ORTHO in the remaining 4/4 pairs. This leave-one-out diagnostic was not a frozen criterion and is not used to rescore the miss.

Unrun controls and certificate denominator. Separate truncation-only, whitening-only, SGD–COPY, and encoder-compatible convolutional controls were not part of the registered design. We draw no component-specific conclusion that would require those controls. The 161 measured cells yield 106 certificates after excluding 41 cells with no defined bound and 14 inside a ± 0.005 abstention band around $\hat{\epsilon}$; re-including that band gives 118 issued and 46 eligible with the same zero violations and the same minimum slack. The certificate audit contains 106 issued certificates, of which 42 pass the recorded eligibility check that the base probe beat the marginal predictor; both denominators are reported.

Payoff-floor sensitivity. The 0.01-bit floor was fixed before E7 and is not tuned. Table 4 re-scores every cross-entropy placebo claim at half and double that floor, using the same rule (paired mean above the floor and positive in every seed) and no new training. No registered AV-MNIST, PCA, or CLIP conclusion changes across this fourfold range. At 0.005 bits one raw-MOSI cell ($+0.009$ bits, 5/5 seeds) would pass. The KT placebo is the only floor-sensitive family: in log-loss it lies at the floor (shared-base KT-1 $+0.0097$, KT-2 $+0.0081$, and CSEDM $+0.0049$ bits; the registered per-arm-base E19 run gives $+0.0109$), which is why KT claims are stated in AUC.

Claim family	Units	0.005	**0.01**	0.02
AV-MNIST image base, rotation (E7)	6 cells	6	**6**	6
AV-MNIST audio base, rotation (E7)	6 cells	6	**6**	6
Other E7 domains incl. raw MOSI/MOSEI, rotation	36 cells	1	**0**	0
Rank-256 PCA, rotation (E8)	6 cells	0	**0**	0
Rank-256 PCA, real modality (E8)	6 cells	6	**6**	6
Frozen CLIP, rotation / copy (E21)	1 / 1	0 / 0	**0 / 0**	0 / 0
KT placebo in log-loss (KT-1, KT-2, CSEDM)	3 cohorts	2	**0**	0

Table 4: Number of units meeting the positive-placebo criterion at three payoff floors (bits); the registered floor is in bold. Computed from the ledger by `floor_sensitivity.py` in the supplement.

B.1 EXPLORATORY CREMA-D SHORT-BUDGET FOLLOW-UP

This follow-up was designed after E7–E24 and is unregistered. We use the released CREMA-D split and the official OGM-GE audio and visual ResNet-18 branches. REAL combines audio and video. COPY, ROT, and GAUSS replace the video branch with an independently trained audio ResNet-18 that receives, respectively, the original spectrogram, a fixed invertible cyclic permutation of it, or fixed per-example Gaussian noise. The fused models differ in trainable parameter count by 0.0281%. All arms use the same 20-epoch budget and five paired seeds; we report epoch 20 rather than selecting each arm's best test epoch.

Gaussian is negative in all five seeds. Thus this experiment is an external example of short-budget optimization relief from decodable information-null branches, not a replication of a positive published real-modality gain and not evidence that rotation generally dominates copy.

Candidate	Accuracy gain (pp)	CE gain (bits)	Positive CE seeds
Real video	-5.054	-0.657	2/5
Copy	-0.806	$+0.122$	5/5
Rotation	$+0.484$	$+0.256$	5/5
Gaussian	-15.538	-1.085	0/5

Table 5: Mean paired final-epoch gains over the audio base on CREMA-D. A positive cross-entropy (CE) gain means lower loss. Rotation's accuracy gain is positive in 3/5 seeds. The information-null controls improve CE, but the criterion's real-gain requirement fails under this short budget.

B.2 Independent MLA late-fusion audit

This unregistered exploratory audit uses the released Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) split and the late-fusion configuration reported by Multimodal Learning with Alternating Unimodal Adaptation (MLA) (Zhang et al., 2024). It is exploratory and is not counted among the registered outcomes. We independently train audio and visual ResNet-18 classifiers and average their frozen logits, without joint fine-tuning. The published configuration reports 66.32% fusion accuracy and 55.48% for its stronger unimodal component. Our gains are measured against the paired visual base; against the stronger unimodal component the copy gain falls to $+1.00$ points and rotation exceeds copy in 2/5 pairs rather than 1/5. Following the released convention, each unimodal checkpoint is selected by its best test accuracy. Base seeds 0–4 are paired with candidate seeds 10000–10004. Copy and rotation use independently trained visual ResNet-18 branches; rotation is a fixed invertible cyclic spatial permutation that preserves convolutional locality.

	Accuracy gain (pp)			CE gain (bits)			Candidate accuracy (%)		
Seed	Real	Copy	Rotation	Real	Copy	Rotation	Real	Copy	Rotation
0	12.500	0.806	-0.134	0.412	0.123	0.084	62.63	63.84	58.06
1	12.903	3.495	-0.269	0.434	0.108	-0.019	62.37	65.46	45.56
2	10.887	4.435	2.151	0.357	0.142	0.129	60.08	66.40	59.81
3	7.392	-2.285	1.075	0.315	0.056	0.137	59.41	62.50	66.53
4	10.349	3.360	0.403	0.305	0.077	0.003	57.66	65.59	53.90
Mean	10.806	1.962	0.645	0.364	0.101	0.067	60.43	64.76	56.77
SD	2.187	2.729	0.994	0.057	0.035	0.071	2.09	1.57	7.75
Positive	5/5	4/5	3/5	5/5	5/5	4/5	—	—	—

Table 6: Five paired seeds from the unregistered MLA audit. Positive cross-entropy (CE) gain means lower loss. Rotation exceeds copy in only 1/5 pairs for either metric; the mean paired rotation-minus-copy differences are -1.317 accuracy points and -0.034 bits (two-sided exact sign-flip $p = .375$ for each). Candidate accuracy is the stand-alone accuracy of the independently trained candidate branch; the rotated branch is weaker and more variable than the copy branch.

The reproduced real-modality gain is large and positive in every seed, whereas both information-null gains are much smaller and copy exceeds rotation on average. The placebo-adjusted gain is 10.161 accuracy points and 0.298 bits, positive in 5/5 pairs. As a sensitivity check on seed 0, excluding the 80 test clips affected by the short-video fallback retains a 9.337-point real-modality gain (75.452% versus 66.114%). The audit therefore distinguishes the published real-modality effect from generic ensemble relief; it does not reproduce that effect with a rotation placebo.

C Experimental setup

C.1 Data and representations

AV-MNIST uses its released image and audio inputs and standard split. MOSI and MOSEI use released aligned text, audio, and vision features. Hateful Memes uses released images and optical-character-recognition (OCR) text; CICIDS2017 (Sharafaldin et al., 2018) uses flow features after fitting all preprocessing on training data. The public KT replication uses the CSEDM 2019 Data Challenge release (Price et al., 2019). Private KT cohorts contain de-identified interaction records and are not released. No test label is used to construct a rotation, PCA transform, or stopping rule. The exploratory CREMA-D follow-up (Cao et al., 2014) uses its released audio and video with

the split distributed by the official On-the-fly Gradient Modulation with Generalization Enhancement (OGM-GE) implementation (Peng et al., 2022). One frame per clip represents video; audio is converted to a log-magnitude short-time Fourier transform using the released loader settings. The independent MLA audit (Appendix B.2) instead follows MLA's released 6,698/744 split and 100-epoch unimodal training budget. Its late-fusion prediction is the arithmetic mean of independently trained, frozen audio and visual logits. We follow the released test-best checkpoint-selection rule and report five fixed base–candidate seed pairs. At one frame per second, 768 of 7,442 released clips yielded fewer than the three frames assumed by the public loader. For these clips we use start, middle, and end source frames, repeating a frame only when necessary; 80 test clips are affected. No audio file required a fallback decoder, and the paper reports a test-set sensitivity analysis that excludes the affected clips.

C.2 Models and pairing

Public-dataset experiments use a gated fusion multilayer perceptron and a two-layer, four-head cross-attention model. Registered budgets are fixed 5 and 20 epochs plus early stopping with a 200-epoch ceiling. Each public configuration uses five paired seeds. The KT comparison uses the same recurrent mixture-of-experts implementation, preprocessing, folds, and stopping rule for the base, real, and rotation arms; only the candidate input changes. Reported differences are paired within split and seed.

For the public dense-input experiments, each gated-fusion branch maps its own input width through 256 and 128 hidden units; the gated and concatenated paths feed a common head (dropout 0.2). Cross-attention projects each modality to 64 dimensions before two four-head layers (dropout 0.1). Both use batches of 256 and AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-3} and weight decay 10^{-4}; the registered SGD comparison uses learning rate 0.05 and momentum 0.9. The base, real, copy, rotation, and Gaussian arms keep these settings fixed. Separate learned input projections handle unequal dense-block widths. This does not establish that an arbitrary structured encoder can decode a flattened rotation: the E24 convolutional case is the counterexample below.

The orthogonal matrix is sampled once per registered cell from a seeded QR matrix factorization and then held fixed. Copy uses the unchanged base. Gaussian uses independent standard-normal columns matched in width and standardized using training statistics. The E8 rank-256 PCA truncation and whitening map is fit on training data only and applied unchanged to validation and test data.

For public-data conditional probes, blocks of at most 48 columns retain their coordinates; wider blocks use 48 train-fitted PCA coordinates. The candidate's observation mask is appended when applicable. In KT, the base block instead combines one-hot current outcome, a ten-bin skill-difficulty indicator, seven learner-history statistics, and out-of-fold base logits and probabilities. The seven statistics are cumulative correctness, within-skill cumulative correctness, log attempt count, five-attempt rolling correctness, cumulative error rate, and the two previous correctness values. These inputs are specified by the frozen probe pipeline; they are not a learned hidden-state summary.

C.3 Metrics and probe gates

We report AUC percentage-point differences where defined and cross-entropy differences in bits. A cross-entropy setting meets the positive-placebo criterion of Section 5 when its paired mean exceeds the fixed 0.01-bit payoff floor and its registered replication threshold is met; task-metric-only audits retain their registered directional threshold. The raw-input headline requires every seed in a cell to agree; E24 registered 5/5 replication. The certificate code applies one eligibility check: the base probe must beat the marginal predictor. Out-of-fold predictions and copy-null convergence are design and diagnostic requirements, not additional filters used to select the reported 42 certificates. A certificate is violated only when the independently trained action exceeds its stated upper bound. We set $\hat{\epsilon} = 0.02\,\mathrm{CE}_p(X)$ in bits, the frozen relative allowance specified before scoring, and $\hat{\delta}$ to the absolute gap between the linear and multilayer-perceptron base probes. Neither slack is tuned to the observed fusion payoff. $\hat{\delta}$ is a heuristic proxy for δ, not a proven upper bound.

For E24, the fixed signed-permutation–discrete-cosine rotation acts on the 784 flattened visual pixels. We reshape its output to 28×28, match the training-audio mean and variance, upsample it to

112×112, and inject it before the official audio ResNet. This preserves zero conditional information given the visual input but not convolutional locality.

C.4 ARTIFACT ORGANIZATION

Measurements are indexed in a single result ledger. The E24 artifact retains five-seed BASE/COPY raw JavaScript Object Notation (JSON) files and all-arm seed-4 JSON; the original seed-0–3 REAL/ROT/GAUSS files were lost with their execution host, so their frozen aggregate is retained instead. The supplement contains the closed experiment plan, raw JSON artifacts for matched KT runs, and the analysis scripts.