BASELINE-RELATIVE PROJECTION DIAGNOSTICS FOR MULTIMODAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fusion gains depend on the nonlinear baseline used for comparison. We develop a reusable diagnostic protocol that applies empirical multimodally-additive projection (EMAP) to both a fusion head and its baseline. Their original gap separates into a projected gap and excess projection sensitivity. The protocol combines shared references, parameter-matched concatenation, documented model selection and paired uncertainty. We characterise finite-reference bias, ranking variance and paired score covariance under explicit assumptions. ANOVA and Shapley relations specify the functional interaction being probed. Experiments combine a historically informed MM-IMDb study with predefined fusion families on MOSI, MOSEI and Balanced Binary Abstract Scenes VQA. On MM-IMDb, Mamba-style retains projected advantages of 1.412 to 1.533 AUC percentage points over parameter-matched concatenation with word2vec features. All projected-gap intervals include zero with sentence-transformer features. On MOSEI, validation selection reduces a source-rate TFN advantage over matched concatenation from 0.753 to -0.001 MAE units. In VQA, MLB-style's original advantage of 2.302 accuracy percentage points becomes a projected gap of -0.005, with interval $[-0.432, 0.401]$. Reference sweeps over 128 to 512 rows distinguish stable VQA interval patterns from reference-sensitive small sentiment effects. An accompanying library implements projection, paired inference and a common reporting record. The protocol makes these properties directly testable components of a fusion comparison.

1 INTRODUCTION

Multimodal fusion heads combine heterogeneous inputs through gating, multiplicative pooling, attention or state-space mixing. Their performance gains are often interpreted through their ability to model cross-modal relationships. A concatenation multilayer perceptron (MLP) can also learn non-additive functions. Consequently, a head's gain can depend on feature processing, optimisation, capacity and the interactions already available to its baseline.

Empirical multimodally-additive projection (EMAP) evaluates a fitted predictor after removing its non-additive component across two input groups (Hessel & Lee, 2020). A useful comparative question follows: how much of a fusion advantage survives when both predictors undergo the same projection? A large response in the head has a different interpretation when concatenation responds similarly. Building on existing projection-based model comparisons (Wenderoth et al., 2025), we organise this report through the original gap G, projected gap G_p and excess projection sensitivity E_p. They satisfy the exact identity $G = G_p + E_p$.

We turn this comparison into a diagnostic protocol with a declared score scale, common reference, capacity controls and paired inference. The estimation analysis describes finite-reference bias and the covariance needed for uncertainty on score contrasts. Established ANOVA and Shapley relations distinguish the functional residual from task-performance sensitivity. The implementation accepts classification or regression predictors and supplies a reusable reporting record.

The empirical study combines historically informed MM-IMDb controls with prospectively specified experiments on MOSI, MOSEI and Balanced Binary Abstract Scenes VQA. The experiments test whether conclusions about a fusion gain persist across stronger baselines, projection and task

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.

metrics. The reporting record contains four original and projected scores, paired contrasts, and the capacity and selection budget.

2 Related Work

Projection and interaction attribution. EMAP introduced performance evaluation after multimodally additive projection, including trained multiplicative controls and reference-subset checks (Hessel & Lee, 2020). Modality Shapley methods characterise input contributions (Hu et al., 2022; Parcalabescu & Frank, 2023). InterSHAP reports EMAP-based performance gaps for both a multimodal variational autoencoder and a nonlinear concatenation baseline (Wenderoth et al., 2025). Multimodal Residual Optimisation separates unimodal and higher-order predictors during learning and evaluates non-additive contributions on MOSI and MOSEI (Wörtwein et al., 2022). Our contribution organises the predictive-score comparison between two fitted models around a shared EMAP reference and paired statistical analysis.

Fusion comparisons. Gating, low-rank bilinear pooling, attention, latent bottlenecks and state-space models provide different fusion mechanisms (Arevalo et al., 2020; Kim et al., 2017; Vaswani et al., 2017; Jaegle et al., 2021; Gu & Dao, 2024). Their performance depends on multimodal optimisation (Wang et al., 2020). Broad comparisons have found competitive late-fusion baselines across datasets (Rheude et al., 2025). MultiBench standardises evaluation across tasks, computational costs and modality robustness (Liang et al., 2021). Our parameter-matched comparisons assess published mechanism families using shared inputs and explicit training budgets. The diagnostic then measures how much of each observed gap remains under a common functional projection.

3 Baseline-Relative Projection Diagnostics

3.1 A shared projection and score convention

Let f output declared model scores and let $p = U|V$ divide the inputs into two groups. Classification uses raw logits and regression uses the scalar prediction. For reference rows $R = \{r_j\}_{j=1}^{n}$, define

$$P_p^R f(x) = \frac{1}{n}\sum_j f(x_U, r_{j,V}) + \frac{1}{n}\sum_j f(r_{j,U}, x_V) - \frac{1}{n^2}\sum_{j,k} f(r_{j,U}, r_{k,V}). \tag{1}$$

The operation is componentwise for vector outputs. It is the least-squares additive projection under the reference product law (Hessel & Lee, 2020). A reference equal to the scoring set recovers the diagonal of the original EMAP grid. A separate reference permits evaluation on new inputs using $2mn + n^2$ cross-pairs for m scoring rows. Cached unimodal transformations reduce repeated encoder computation.

Figure 1 connects this construction to the four-score comparison.

Let S be a specified utility for which larger values are better. We use macro-ROC-AUC, negative mean absolute error (MAE), or official VQA accuracy. The *own-model projection sensitivity* is

$$D_p(f) = S(f) - S(P_p^R f). \tag{2}$$

A positive sensitivity indicates lower task performance after projection. Its value depends on the score scale, metric and reference. For three input blocks a, b, c, the partitions $ab|c$, $ac|b$ and $a|bc$ retain relationships within each grouped side. For example, $P_{ab|c}^R f$ retains interactions between a and b. Each partition therefore defines a distinct diagnostic.

3.2 Paired contrasts and a reusable procedure

For a fusion head f_h and baseline f_c, define

$$G = S(f_h) - S(f_c), \qquad G_p = S(P_p^R f_h) - S(P_p^R f_c), \tag{3}$$

$$E_p = D_p(f_h) - D_p(f_c), \qquad G = G_p + E_p. \tag{4}$$

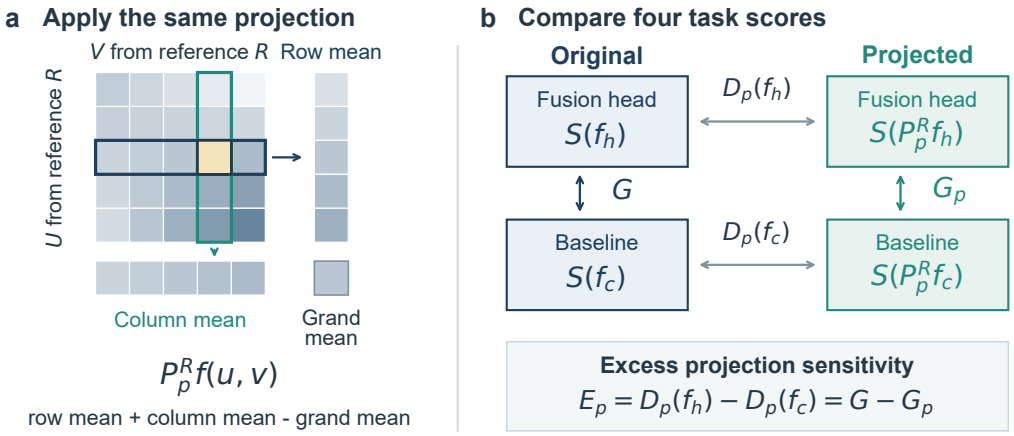

Figure 1: Shared projection and the four-score comparison. (a) An illustrative reference grid shows the row mean plus column mean minus grand mean construction. (b) Applying this operation to both predictors yields four task utilities. Horizontal contrasts measure each model's projection sensitivity. Vertical contrasts measure the original and projected gaps. Their difference gives excess projection sensitivity E_p. In (b), colours identify original and projected scores.

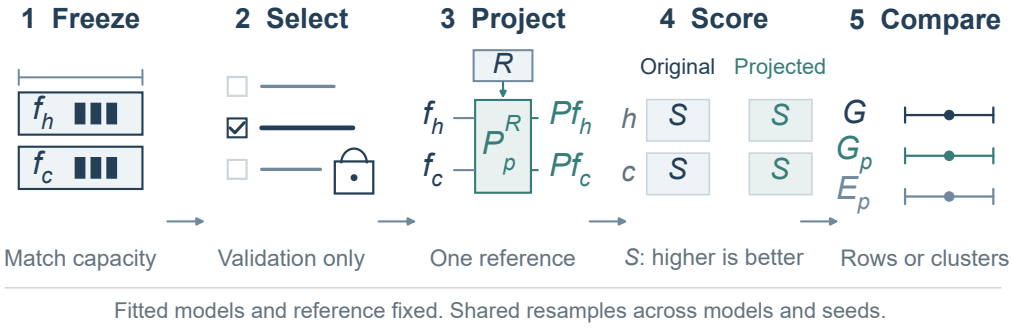

Figure 2: Diagnostic workflow. A common reference and paired resampling preserve the dependence among the four measured scores. Here $P = P_p^R$. Icons and intervals illustrate the procedure.

The projected gap G_p measures the advantage remaining after the shared projection. Excess projection sensitivity E_p compares the models' own sensitivities. These signed contrasts remain interpretable when G is near zero. Each contrast is computed within training seed before the seed mean is taken.

Using the diagnostic. Model selection starts from validation utility and the deployment resource budget. Applications that value projected performance can specify $U_\lambda(f) = (1-\lambda)S(f) + \lambda S(P_p^R f)$, with $0 \le \lambda \le 1$. Under a common baseline, this utility gives the advantage $G - \lambda E_p$. At equal G, smaller E_p gives greater combined utility for $\lambda > 0$. The choice of λ precedes validation selection and identifies the evaluation objective. Deployment with missing modalities or distribution shift requires evaluation under the corresponding conditions. Section B gives a worked example and a linear-size partition budget for many input blocks.

For independent Rademacher inputs, let $y = \mathbf{1}\{ab = 1\}$ and let both predictors equal ab. Under the population product law, each achieves AUC 1 and projects to a constant with AUC 0.5. Both have sensitivity 0.5, while $G = G_p = E_p = 0$. Their shared interaction supports prediction equally well. Reporting both sensitivities alongside the three gaps makes this pattern visible.

The procedure in Figure 2 takes fitted predictors, scoring rows, a reference and a metric. It returns four scores, G, G_p, E_p, both D_p values and pointwise intervals. Each bootstrap draw is shared across

models, partitions and fitted seeds. For clustered data, a draw retains every row from each sampled source unit. The accompanying library supports exact and sampled reference integration, custom metrics and a reporting schema. The schema records features, splits, trainable parameters, selection budgets, reference construction, sampling units and the conditioning of uncertainty. Complete pseudocode appears in Figure 4.

3.3 Statistical and functional interpretation

We specialise established functional-decomposition and correlated-score inference results to the shared-reference comparison. Under a product law $Q = J_U \otimes J_V$, the residual $r = f - P_Q f$ is the two-group functional ANOVA interaction. For the corresponding two-player marginal-reference game, r equals twice the off-diagonal SHAP interaction value (Owen, 2012; Lundberg et al., 2018). With a common fully factorised three-block law, $f - P_{ab|c} f = f_{ac} + f_{bc} + f_{abc}$ and $\Phi_{ac} = f_{ac}/2 + f_{abc}/4$. Dependent grouped references require separate two-group interpretations. Section A also distinguishes score sensitivity from interaction information.

For n independent reference pairs from J, independent of fitting and scoring, finite expectations give the pointwise bias

$$\mathbb{E}[\widehat{P}_n f(x)] - P_Q f(x) = -\frac{\mathbb{E}_J f - \mathbb{E}_Q f}{n}. \tag{5}$$

An off-diagonal grand mean removes this centring bias. Its correction is constant across inputs for each output, leaving each label's AUC unchanged. Unequal output-specific offsets can change an argmax decision, including binary VQA. A scalar offset can change MAE. For two fixed scoring inputs, the estimated projected margin is an average of $W_f(Z) = f(u, V) - f(u', V) + f(U, v) - f(U, v')$. If W_f has a finite second moment, its variance is $\mathrm{Var}_J(W_f)/n$. Shared reference rows induce cross-model covariance of the same order. Under bounded scores, a margin-based concentration bound quantifies the remaining reference-induced AUC bias.

Let θ contain the four population utilities and $\widehat{\theta}$ their scoring-sample estimates. A linear contrast matrix L gives $(G, G_p, E_p, D_p(f_h), D_p(f_c))^\top = L\theta$. Thus

$$\begin{aligned} \mathrm{Cov}(L\widehat{\theta}) &= L\,\mathrm{Cov}(\widehat{\theta})L^\top, \\ \mathrm{Var}(\widehat{E}_p) &= \mathrm{Var}(\widehat{G}) + \mathrm{Var}(\widehat{G}_p) - 2\,\mathrm{Cov}(\widehat{G}, \widehat{G}_p). \end{aligned} \tag{6}$$

For identically distributed independent scoring rows and negative MAE, the first covariance equals $L\Sigma L^\top/m$, where Σ is the covariance of the four per-row utilities. AUC admits a two-sample U-statistic decomposition (Hoeffding, 1948; DeLong et al., 1988). Our video and scene experiments preserve source clustering in resampling. The full assumptions, proofs and finite-support checks are in Section A. Separate-reference intervals estimate scoring uncertainty conditional on fitted models and reference. The full-test movie intervals describe stability under row reweighting. Training-row references support the conditional scoring analysis, while the reference-bias and concentration results require independence from fitting.

4 Experimental Design

Parameter-matched concatenation matches each fusion head's trainable capacity. Shared concatenation uses one baseline architecture across heads within a task.

Movie genre classification. MM-IMDb provides 15,552 training, 2,608 validation and 7,799 test movies with 23 genre labels (Arevalo et al., 2020). Our three-block extension uses plot, poster and production metadata as a, b, c. The word2vec bank (w2v) has 300-dimensional text features. The sentence-transformer bank (ST) has 768-dimensional `all-mpnet-base-v2` features (Reimers & Gurevych, 2019). Both use the same 1,024-dimensional DenseNet-121 poster features (Huang et al., 2017), 44 metadata features and movie splits. Feature standardisation and metadata category selection use training data. Metadata excludes audience ratings and vote counts. Macro-AUC averages all 23 genre AUCs.

Movie models and capacity controls. The Mamba-style head uses three pooled tokens, width 256, state size 16 and two mixing blocks per direction. Concatenation uses one hidden ReLU layer.

Strictly additive prediction sums independent width-256 logit branches. All use AdamW and binary cross-entropy for 30 epochs with seeds 0, 1 and 2. Validation macro-AUC selects checkpoints. The capacity study searches three learning rates at concatenation widths 256, 512, 1,024 and the nearest parameter-matched width. The capacity study contains 72 fit records and eight width-specific selections. Mamba-style led the earlier movie-head comparison and was selected for this analysis. Sections D and J specify the full recipes and selection records.

Audiovisual-language sentiment regression. We use the MMSA aligned releases of MOSI and MOSEI (Zadeh et al., 2016; Bagher Zadeh et al., 2018; THUIAR, 2025). Their training/validation/test splits contain 1,284/229/686 and 16,326/1,871/4,659 segments. Each segment has 50-step text, audio and visual features, with dimensions 768/5/20 and 768/74/35, respectively. LMF and TFN use the MMSA reference implementations (Liu et al., 2018; Zadeh et al., 2017). Text passes through an LSTM, while audio and visual features are averaged over their 50 steps before their subnetworks. Each matched concatenation model shares the corresponding unimodal architectures and initial weights, then trains independently with a matched readout. Trainable counts differ by at most two parameters. Shared concatenation averages each modality over time and uses an MLP with widths 256 and 128.

All models use Adam, MAE loss, at most 60 epochs, validation patience 8 and gradient clipping at 5. Each model receives three learning-rate candidates and seeds 0, 1 and 2. The candidates are the source rate plus 10^{-4} and 3×10^{-4}. The source rate is 5×10^{-3} for MOSEI TFN and its matched concatenation, and 10^{-3} for the other models. Validation MAE selects each checkpoint and its three-seed mean selects the learning rate. The utility is negative MAE, with partitions $T|AV$, $A|TV$ and $V|TA$. The test splits contain 31 MOSI videos and 676 MOSEI videos. Section F specifies batch sizes, regularisation and all model definitions.

Balanced Binary Abstract Scenes VQA. We use the official RGB question-answering task of Zhang et al. (2016); Antol et al. (2015), with 22,055 training and 11,328 held-out validation questions. A frozen ResNet-18 supplies 512-dimensional image features (He et al., 2016), alongside 4,096 question TF-IDF features fitted on internal training only. The internal split keeps source scenes and their complementary scenes together, giving 18,745 training and 3,310 validation questions. We predefine MLB-style and GMU-style heads, shared width-128 concatenation and parameter-matched concatenation for each head (Kim et al., 2017; Arevalo et al., 2017). MLB-style uses width-256 tanh projections and a Hadamard product. GMU-style uses width-256 projections and a sigmoid gate. Matched concatenation widths are 256 and 512, with parameter differences of 0.0217% and 0.0108%. All five models receive learning rates 3×10^{-4} and 10^{-3} and three seeds. They train for 30 epochs with AdamW, batch size 256 and weight decay 10^{-4}. Binary cross-entropy fits two answer logits to the official VQA-credit targets. Internal validation accuracy selects checkpoints and the three-seed mean selects the learning rate. The diagnostic uses official ten-answer VQA accuracy and projects both raw logits before the yes/no argmax decision. The held-out scoring set has 5,368 source-scene groups.

Prospective selection and inference. The sentiment and VQA model rosters are fixed before initial fitting. A pre-test sentiment amendment adds the same two smaller learning rates to every model after examining validation optimisation. The pre-scoring records seal 72 capacity-study fits and eight choices, 90 sentiment fits and ten choices, and 30 VQA fits and five choices. Local protocol records and artifact hashes document this sequence. All declared fusion families and seeds are retained in the results. The primary new-task comparisons share 256 training-reference rows and exact cross-product integration within each dataset. Test observations remain separate from reference construction. Shared bootstrap draws resample source videos for sentiment and source scenes for VQA. Movie intervals use 1,000 paired row draws. The main movie capacity comparison keeps the complete test reference fixed, and a separate training-reference analysis assesses that design choice. All intervals are pointwise 95% percentile intervals conditional on fitted models, configurations, seeds and reference. Sentiment uses 2,000 replicates and VQA uses 1,000.

5 RESULTS

5.1 MECHANISM CONTROLS

Three finite-support tasks use additive, product and mixed label mechanisms. Additive, concatenation and bilinear-plus-additive predictors give 27 validation-selected fits. Table 1 reports concatenation across all three mechanisms. Exact coalition enumeration gives group-SHAP interaction RMS equal to half the residual RMS under the same two-group product reference. The product task has smaller residual RMS than the mixed task but larger AUC sensitivity. Both nonlinear models have mixed-task $D_p = 21.897$ AUC percentage points, yielding paired $E_p \approx 0$ and $E_p = 21.897$ against the additive model. Functional interaction magnitude and task sensitivity therefore need not vary together. Sections O and O.1 report every model and seed, with analytic controls.

Table 1: Concatenation MLPs across all three label mechanisms. Residual and group-SHAP RMS use raw logits. D_p uses AUC percentage points (pp). Entries are three-seed means $\pm$ sample SD. Appendix O.1 reports all nine conditions.

Mechanism	Residual RMS	Group-SHAP RMS	D_p (pp)
Additive	0.168 ± 0.064	0.084 ± 0.032	-0.052 ± 0.090
Product	0.691 ± 0.163	0.346 ± 0.081	50.187 ± 0.186
Mixed	1.683 ± 0.706	0.842 ± 0.353	21.897 ± 0.000

5.2 BASELINE STRENGTH AND PARAMETER MATCHING ON MM-IMDB

Mamba-style reaches mean macro-AUC 0.8710 on w2v and 0.9344 on ST. Validation-selected concatenation reaches 0.8567 and 0.9332. On ST, selection reduces the original Mamba-style gap from 0.293 to 0.115 AUC percentage points (pp), with interval $[-0.012, 0.239]$. The three projected gaps against selected concatenation also have intervals spanning zero. On w2v, projected gaps remain 1.355 to 1.421 pp, each with an interval above zero. Original scores for all four models and their paired contrasts appear in Tables 43 and 44.

Parameter matching uses concatenation widths 1,658 on w2v and 1,305 on ST. Every added parameter participates in prediction. The resulting counts differ from Mamba-style by 0.0069% and 0.0249%. Matched concatenation reaches AUC 0.8559 and 0.9335. Across the four tested widths, w2v concatenation ranges from 0.8557 to 0.8567, while ST ranges from 0.9328 to 0.9335. Extra capacity therefore has different effects across the two representations.

Table 2: Mamba-style versus parameter-matched concatenation on MM-IMDb. Values are three-seed means with pointwise 95% paired resampling intervals, in AUC percentage points. The full-test projection reference remains fixed during resampling.

Text	Partition	G	G_p	E_p
w2v	$ab \mid c$	1.504 $[1.236, 1.765]$	1.412 $[1.146, 1.659]$	0.092 $[0.024, 0.170]$
w2v	$ac \mid b$	1.504 $[1.236, 1.765]$	1.465 $[1.195, 1.721]$	0.040 $[-0.039, 0.127]$
w2v	$a \mid bc$	1.504 $[1.236, 1.765]$	1.533 $[1.273, 1.776]$	-0.029 $[-0.116, 0.063]$
ST	$ab \mid c$	0.086 $[-0.030, 0.212]$	-0.025 $[-0.144, 0.099]$	0.111 $[0.071, 0.152]$
ST	$ac \mid b$	0.086 $[-0.030, 0.212]$	0.064 $[-0.057, 0.187]$	0.022 $[-0.019, 0.062]$
ST	$a \mid bc$	0.086 $[-0.030, 0.212]$	0.028 $[-0.078, 0.149]$	0.058 $[0.008, 0.106]$

On w2v, all projected advantages over matched concatenation remain positive, from 1.412 to 1.533 pp (Table 2). Excess sensitivity is positive under $ab|c$, where the head and baseline have own-model sensitivities of 0.289 and 0.197 pp. Its other two excess-sensitivity intervals span zero. On ST, the

original gap is 0.086 pp with interval $[-0.030, 0.212]$, and all three projected-gap intervals span zero. Excess sensitivity is positive under $ab|c$ and $a|bc$. Under $ab|c$, the head and baseline sensitivities are 0.181 and 0.070 pp. The joint pattern distinguishes a persistent projected advantage on w2v from uncertain small gaps on ST. Section D gives the full capacity curve and own-model sensitivities.

Training-reference capacity check. A secondary analysis retained the capacity-matched models and used 256 training movies as the projection reference. On w2v, projected gaps ranged from 1.413 to 1.524 AUC percentage points. All three projected-gap intervals were positive. On ST, projected gaps ranged from -0.030 to 0.060 points. All three projected-gap intervals included zero. The full-test-reference comparison remains the primary capacity analysis. Section E gives the secondary protocol and complete scores.

5.3 Capacity and baseline strength in sentiment regression

On MOSI, LMF has MAE 1.126 ± 0.290 against matched concatenation's 1.292 ± 0.309. TFN scores 0.944 ± 0.010 against 0.978 ± 0.040. Values are means and sample SDs across three seeds. Shared concatenation reaches 0.949 ± 0.024. LMF's original gap changes from 0.166 against matched concatenation to -0.177 against shared concatenation, with interval $[-0.224, -0.129]$. TFN's gap against shared concatenation is 0.005, with interval $[-0.034, 0.041]$. The shared comparison therefore measures baseline predictive strength alongside parameter matching.

On MOSEI, LMF scores 0.565 ± 0.005 against matched concatenation's 0.665 ± 0.150 and shared concatenation's 0.576 ± 0.003. Its original advantage over shared concatenation is 0.011, with interval $[0.003, 0.019]$. TFN scores 0.5783 ± 0.0092 against matched concatenation's 0.5775 ± 0.0027. Validation selection changes this control's learning rate from 0.005 to 0.0003 and lowers its MAE from 1.331 ± 1.322 to 0.578 ± 0.003. The original TFN gap consequently falls from 0.753 under the source-rate recipe to -0.001, with interval $[-0.008, 0.006]$. The selected TFN gap against shared concatenation is -0.002, with interval $[-0.010, 0.006]$. Section F reports the complete source-rate analysis, validation choices and per-seed behaviour.

At the primary 256-row reference, projected LMF intervals remain positive against matched concatenation on both datasets (Table 3). MOSI TFN retains positive projected-gap intervals under $T|AV$ and $V|TA$, while the $A|TV$ interval spans zero. All three projected TFN gaps on MOSEI have intervals spanning zero. For MOSI TFN under $V|TA$, the head and matched concatenation have own-model sensitivities of -0.0101 and -0.0011. Both mean MAEs decrease after projection, yielding $E_p = -0.0090$ with interval $[-0.0180, -0.0008]$.

5.4 Distinct projection responses in visual question answering

MLB-style achieves 67.10% original accuracy against matched concatenation's 64.80%. With the primary 256-row reference, both projected models score 64.57%. Its original advantage of 2.302 pp becomes a projected gap of -0.005 pp, with interval $[-0.432, 0.401]$ (Table 4). The head's own sensitivity is 2.536 pp and matched concatenation's is 0.229 pp, yielding excess sensitivity 2.306 pp. This comparison associates the observed original advantage with differential response to the specified projection.

GMU-style shows a different pattern. Original accuracy is 65.34% against matched concatenation's 64.89%, and the original-gap interval spans zero. Projected accuracy is 65.25% against 64.53%, giving $G_p = 0.716$ pp with interval $[0.236, 1.170]$. The baseline loses more accuracy under projection, so $E_p = -0.268$ pp. Relative to shared concatenation, both its original and projected gap intervals are above zero. The sign compares task sensitivities under the specified projection and metric. Functional additivity is assessed separately through each model's residual function.

Figure 3 summarises three ways the measured comparison changes: baseline choice, validation-based learning-rate selection and shared projection.

5.5 Reference sensitivity and computational cost

The reference study evaluates all selected sentiment and VQA models. Each dataset uses 18 conditions: three draws at $n \in \{128, 256, 512\}$ under observation-uniform and source-uniform training-

Table 3: Sentiment models against matched concatenation after validation selection. Values are three-seed means and pointwise 95% source-video bootstrap intervals in MAE units. Positive G or G_p favours the fusion model. Intervals condition on the fitted models and training reference.

Task / head	Partition	G	G_p	E_p
MOSI / LMF	$T \mid AV$	0.1657 $[0.1235, 0.2109]$	0.1695 $[0.1271, 0.2148]$	-0.0038 $[-0.0107, 0.0023]$
	$A \mid TV$	0.1657 $[0.1235, 0.2109]$	0.1660 $[0.1240, 0.2108]$	-0.0003 $[-0.0024, 0.0016]$
	$V \mid TA$	0.1657 $[0.1235, 0.2109]$	0.1697 $[0.1272, 0.2155]$	-0.0039 $[-0.0099, 0.0014]$
MOSI / TFN	$T \mid AV$	0.0345 $[0.0007, 0.0696]$	0.0410 $[0.0106, 0.0726]$	-0.0065 $[-0.0164, 0.0024]$
	$A \mid TV$	0.0345 $[0.0007, 0.0696]$	0.0328 $[-0.0005, 0.0677]$	0.0017 $[-0.0024, 0.0054]$
	$V \mid TA$	0.0345 $[0.0007, 0.0696]$	0.0435 $[0.0121, 0.0752]$	-0.0090 $[-0.0180, -0.0008]$
MOSEI / LMF	$T \mid AV$	0.1005 $[0.0896, 0.1119]$	0.1006 $[0.0894, 0.1118]$	-0.0001 $[-0.0022, 0.0020]$
	$A \mid TV$	0.1005 $[0.0896, 0.1119]$	0.1008 $[0.0898, 0.1120]$	-0.0003 $[-0.0009, 0.0004]$
	$V \mid TA$	0.1005 $[0.0896, 0.1119]$	0.1005 $[0.0895, 0.1117]$	0.0001 $[-0.0019, 0.0020]$
MOSEI / TFN	$T \mid AV$	-0.0008 $[-0.0077, 0.0061]$	-0.0005 $[-0.0072, 0.0061]$	-0.0003 $[-0.0019, 0.0013]$
	$A \mid TV$	-0.0008 $[-0.0077, 0.0061]$	-0.0005 $[-0.0074, 0.0063]$	-0.0003 $[-0.0014, 0.0008]$
	$V \mid TA$	-0.0008 $[-0.0077, 0.0061]$	-0.0009 $[-0.0078, 0.0060]$	0.0001 $[-0.0012, 0.0014]$

Table 4: Balanced Binary Abstract Scenes VQA. Three-seed mean contrasts and pointwise 95% source-scene bootstrap intervals, in accuracy percentage points. Shared concatenation has width 128. Matched widths are 256 for MLB-style and 512 for GMU-style. Contrasts are computed before rounding.

Head / baseline	G	G_p	E_p
MLB-style / shared	2.208 $[1.658, 2.773]$	-0.195 $[-0.636, 0.247]$	2.403 $[1.836, 2.972]$
MLB-style / matched	2.302 $[1.732, 2.831]$	-0.005 $[-0.432, 0.401]$	2.306 $[1.729, 2.867]$
GMU-style / shared	0.446 $[0.116, 0.789]$	0.487 $[0.150, 0.810]$	-0.041 $[-0.191, 0.118]$
GMU-style / matched	0.448 $[-0.043, 0.871]$	0.716 $[0.236, 1.170]$	-0.268 $[-0.445, -0.084]$

reference laws. The latter weights videos or scenes equally before sampling an observation within each source. This changes the reference population, while repeated draws measure sampling variation within each law. Each condition retains the full scoring set and shared cluster-bootstrap draws.

For VQA MLB-style against matched concatenation, G_p ranges from -0.005 to 0.277 pp, with all 18 intervals spanning zero. Its E_p ranges from 2.025 to 2.306 pp, with all intervals positive. Against matched concatenation, GMU-style retains positive G_p intervals and negative E_p intervals in all 18 conditions. Against matched concatenation on MOSI, TFN's $V|TA$ excess sensitivity ranges from -0.00904 to -0.00740 MAE units. Only three of its 18 intervals exclude zero. Under $A|TV$, one TFN projected-gap interval is positive and 17 span zero. These small effects require reference-specific interpretation. LMF retains positive G_p intervals against matched concatenation under every partition and reference on both sentiment datasets. For MOSEI TFN against matched concatenation, all G_p and E_p intervals span zero across the three partitions and 18 conditions. Excess-sensitivity

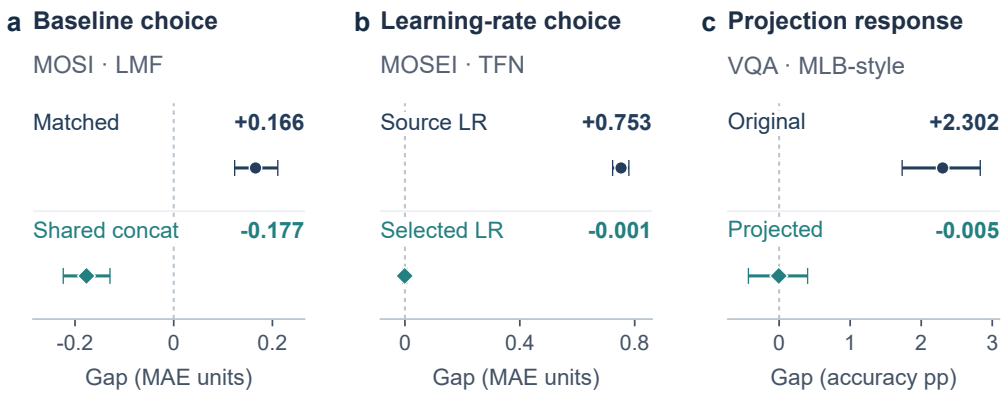

Figure 3: Three diagnostic contrasts. (a) LMF on MOSI against matched and shared concatenation. (b) TFN on MOSEI before and after validation-based learning-rate selection. (c) MLB-style on Balanced Binary Abstract Scenes VQA before and after projection, against matched concatenation. Panels (a,b) show G in MAE units. Panel (c) shows G and G_p in accuracy percentage points. Points average three fitted seeds. Bars are pointwise 95% paired cluster-bootstrap intervals conditional on fixed fits and the 256-row training reference. Resampling units are 31 MOSI videos, 676 MOSEI videos and 5,368 VQA source scenes. Vertical lines mark zero.

intervals against shared concatenation also change between excluding and spanning zero in several sentiment comparisons. Section H reports all partitions, reference sizes and laws, separating draw variation from conditional scoring intervals.

Exact integration uses $2mn + n^2$ cross-pairs per model and partition. For VQA, the incremental $n = 256$ diagnostic takes 0.20–0.31 seconds per model, 39.5–79.1 times ordinary scoring from stored features. Section I separates preparation, scoring and projection costs and documents the timing conditions.

6 Discussion

First, fusion advantages depend on the baseline and its optimisation. Selected concatenation narrows the ST Mamba-style gap from 0.293 to 0.115 AUC pp. The MOSI LMF gap changes sign between matched and shared concatenation (Figure 3a). Learning-rate selection reduces the MOSEI TFN gap from 0.753 to -0.001 MAE units (Figure 3b).

Second, projection responses differ across models. MLB-style's original VQA gap nearly equals its excess sensitivity, leaving G_p near zero (Figure 3c). GMU-style retains positive G_p because matched concatenation loses more accuracy under projection. Sensitivity measures the utility change on removing the reference-specific non-additive component.

Third, VQA interval patterns and LMF projected-gap intervals persist across all 18 reference conditions. Small sentiment contrasts, including TFN excess sensitivity under $V|TA$, move between excluding and spanning zero.

Cross-pairing can form combinations rare under the observed joint distribution. Results are interpreted relative to the declared reference law.

7 Conclusion

Shared-reference projection expresses a fusion comparison as four scores and three paired contrasts. Parameter matching, documented selection and resampling by source unit make the comparison controlled and auditable. The theory identifies the functional interactions and covariance assumptions behind inference. The protocol tests whether fusion gains survive stronger baselines and additive projection across multimodal tasks.

Reproducibility Statement

The anonymous release includes the diagnostic library, reporting schema, frozen protocols, model implementations, aggregate results and numerical checks. New sentiment and VQA experiments provide source hashes and executable acquisition and training procedures for public datasets. Movie experiments use prepared feature banks and recorded splits. Movie protocols, feature construction procedures and feature-bank hashes appear in the appendix.

Ethics Statement

The experiments use public benchmark releases under their providers' research terms. Cross-paired inputs serve as computational probes of the fitted models. The release supplies code and aggregate evidence, with real-benchmark examples and model weights retained separately.

Disclosure of Large Language Model Use

AI tools supported method development and evaluation, as well as manuscript refinement. Empirical results were verified against saved outputs, and the authors take responsibility for the final work.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. URL https://arxiv.org/abs/1505.00468.

John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A. Gonzalez. Gated multimodal units for information fusion. In *ICLR Workshop*, 2017. URL https://arxiv.org/abs/1702.01992.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal networks. *Neural Computing and Applications*, 32(14):10209–10228, 2020. doi: 10.1007/s00521-019-04559-1.

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988. doi: 10.2307/2531595. URL https://pubmed.ncbi.nlm.nih.gov/3203132/.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

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. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

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. doi: 10.18653/v1/2020.emnlp-main.62.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Pengbo Hu, Xingyu Li, and Yi Zhou. SHAPE: An unified approach to evaluate the contribution and cooperation of individual modalities. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3064–3070, 2022. doi: 10.24963/ijcai.2022/425.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, pp. 4651–4664, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=r1rhWnZkg.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. MultiBench: Multiscale benchmarks for multimodal representation learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/37693cfc748049e45d87b8c7d8b9aacd-Abstract-round1.html.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2256, 2018. doi: 10.18653/v1/P18-1209. URL https://aclanthology.org/P18-1209/.

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles, 2018. URL https://arxiv.org/abs/1802.03888.

William J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, 1954. doi: 10.1007/BF02289159. URL https://www.cambridge.org/core/journals/psychometrika/article/abs/multivariate-information-transmission/7EB2980C5C5668747D306E2842C8C0A8.

Art B. Owen. Variance components and generalized Sobol' indices, 2012. URL https://arxiv.org/abs/1205.1774.

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)*, pp. 4032–4059, 2023. doi: 10.18653/v1/2023.acl-long.223.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Tillmann Rheude, Roland Eils, and Benjamin Wild. Fusion or confusion? multimodal complexity is not all you need. *arXiv preprint arXiv:2512.22991*, 2025. doi: 10.48550/arXiv.2512.22991. URL https://arxiv.org/abs/2512.22991v1.

THUIAR. MMSA: A unified framework for multimodal sentiment analysis. GitHub repository, 2025. URL https://github.com/thuiar/MMSA/tree/a94e65d07fa1ae0d44e552390074b29b0898edfd. Code and feature manifest at commit a94e65d07fa1ae0d44e552390074b29b0898edfd, accessed 12 September 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705, 2020. doi: 10.1109/CVPR42600.2020.01271.

Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21501–21509, 2025. doi: 10.1609/aaai.v39i20.35452. URL https://ojs.aaai.org/index.php/AAAI/article/view/35452.

Torsten Wörtwein, Lisa Sheeber, Nicholas Allen, Jeffrey Cohn, and Louis-Philippe Morency. Beyond additive fusion: Learning non-additive multimodal interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4681–4696, 2022. doi: 10.18653/v1/2022.findings-emnlp.344.

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. doi: 10.48550/arXiv.1606.06259. URL https://arxiv.org/abs/1606.06259.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115/.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. URL https://arxiv.org/abs/1511.05099.

A Reference estimation and paired score inference

We distinguish the reference law from the labelled evaluation law. Let J describe reference pairs (U, V) and set $Q = J_U \otimes J_V$. Let P describe evaluation observations (X, Y), where $X = (U, V)$. The AUC results assume $Y \in \{0, 1\}$. The regression results allow real-valued labels. The fitted predictor is fixed throughout this analysis. The independent-reference moment and concentration results additionally assume independence from training and model selection. Reference rows used during training retain their dependence on the fitted model. For a scalar output $f \in L^2(Q)$, define

$$A_f(u) = \mathbb{E}_{J_V} f(u, V), \qquad B_f(v) = \mathbb{E}_{J_U} f(U, v), \qquad \mu_Q = \mathbb{E}_Q f.$$

The population projection is $P_Q f(u, v) = A_f(u) + B_f(v) - \mu_Q$. All evaluations below use versions defined on the required cross-pairs. The AUC bounds additionally require bounded logits on those pairs.

Proposition 1 (Projection geometry and interaction attribution)**.** *The map P_Q is the orthogonal projection onto the closed additive subspace $\{c + g(U) + h(V) : \mathbb{E}g = \mathbb{E}h = 0\} \subset L^2(Q)$. Its residual $r = f - P_Q f$ satisfies*

$$\mathbb{E}_Q[r \mid U] = \mathbb{E}_Q[r \mid V] = 0, \qquad \|f\|_Q^2 = \|P_Q f\|_Q^2 + \|r\|_Q^2.$$

For the two-player marginal-reference game $v_x(S) = \mathbb{E}_Q[f(x_S, X_{-S})]$, the mixed game difference equals $r(x)$. The off-diagonal SHAP interaction value equals $r(x)/2$ under the symmetric matrix convention.

Proof. Under Q, independence gives $\mathbb{E}[B_f(V) \mid U] = \mu_Q$. Subtracting $P_Q f$ therefore gives $\mathbb{E}_Q[r \mid U] = 0$. The corresponding calculation conditions on V. Thus r is orthogonal to every additive function, while $P_Q f$ is additive. The constant, centred U and centred V subspaces are mutually orthogonal and closed. Their finite orthogonal sum is closed, establishing the projection and Pythagorean identities. Finally, the game difference is $v_x(UV) - v_x(U) - v_x(V) + v_x(\varnothing) = f(x) - A_f(u) - B_f(v) + \mu_Q$. The off-diagonal SHAP convention assigns half this difference to each matrix entry. □

This proposition specialises functional ANOVA and the SHAP interaction definition (Owen, 2012; Lundberg et al., 2018) to the EMAP reference (Hessel & Lee, 2020). For a common fully factorised three-block reference, write the ANOVA components as f_T. Then

$$f - P_{ab|c} f = f_{ac} + f_{bc} + f_{abc}, \qquad \Phi_{ac} = \tfrac{1}{2} f_{ac} + \tfrac{1}{4} f_{abc}.$$

Each bisection thus retains within-group components and removes components crossing its boundary. With dependent blocks, $J_{ab} \otimes J_c$ can differ from $J_a \otimes J_b \otimes J_c$. The two-group result remains valid for each specified reference. A shared three-block ANOVA interpretation requires the common factorised reference. Conditional-expectation SHAP under the observed dependent law defines a different game.

Proposition 2 (Reference bias and ranking variance)**.** *Let $Z_i = (U_i, V_i)$ be n independent draws from J, independent of the evaluation observations. Assume that their distribution remains J^n conditional on the fitted predictor. For fixed $x = (u, v)$, define*

$$\widehat{A}_f(u) = \frac{1}{n} \sum_i f(u, V_i), \quad \widehat{B}_f(v) = \frac{1}{n} \sum_i f(U_i, v), \quad \widehat{\mu}_V = \frac{1}{n^2} \sum_{i,j} f(U_i, V_j).$$

If the displayed expectations exist and $\mu_J = \mathbb{E}_J f$, then

$$\mathbb{E}[\widehat{P}_n f(x)] - P_Q f(x) = -\frac{\mu_J - \mu_Q}{n}, \qquad \widehat{P}_n f = \widehat{A}_f + \widehat{B}_f - \widehat{\mu}_V. \tag{7}$$

For $n \geq 2$, replacing $\widehat{\mu}_V$ by $\widehat{\mu}_U = [n(n-1)]^{-1} \sum_{i \neq j} f(U_i, V_j)$ removes this pointwise bias. Both centring choices give exactly the same ranking.

For fixed evaluation inputs x, x', set

$$W_f(Z; x, x') = f(u, V) - f(u', V) + f(U, v) - f(U, v').$$

The projected margin $\widehat{\Delta}_f = \widehat{P}_n f(x) - \widehat{P}_n f(x')$ obeys the following moment identities whenever $\mathbb{E}_J W_f^2 < \infty$:

$$\mathbb{E}\widehat{\Delta}_f = P_Q f(x) - P_Q f(x'), \qquad \mathrm{Var}(\widehat{\Delta}_f) = \frac{\mathrm{Var}_J W_f}{n}.$$

When two models share the reference and $\mathbb{E}_J W_g^2 < \infty$, their margin covariance is $\mathrm{Cov}(\widehat{\Delta}_f, \widehat{\Delta}_g) = \mathrm{Cov}_J(W_f, W_g)/n$.

Proof. The n diagonal terms in $\widehat{\mu}_V$ have expectation μ_J. The $n(n-1)$ off-diagonal terms have expectation μ_Q. Both marginal means are unbiased at fixed evaluation inputs. This proves the bias formula and the off-diagonal correction. The estimated grand mean is constant across evaluation inputs, so it cancels from every score difference. The remaining difference is $n^{-1} \sum_i W_f(Z_i; x, x')$. Independence across reference rows gives its expectation, variance and cross-model covariance. □

The grand mean is a V-statistic, while its off-diagonal correction is a symmetrised U-statistic (Hoeffding, 1948). For AUC alone, $\widehat{A}_f + \widehat{B}_f$ gives exactly the centred projection's ranking. This saves the reference-by-reference grid when scoring an independent sample. For m evaluation inputs, the remaining computation uses $2mn$ cross-pair evaluations per model and partition. Centred logits remain necessary for quantities that depend on absolute score values. For negative mean absolute error, the grand mean must be retained. Changing V-statistic centring to U-statistic centring changes that finite-reference metric and its estimand. The moment calculation describes both definitions without selecting between them.

Proposition 3 (Reference error and ranking sensitivity)**.** *Use the independent-reference setup of Proposition 2. Assume $|f| \le B$ on the relevant evaluation-reference cross-pairs, with $B > 0$. Let $X^+ \sim P(X \mid Y = 1)$ and $X^- \sim P(X \mid Y = 0)$ be mutually independent and independent of the reference. Write $\Delta_f = P_Q f(X^+) - P_Q f(X^-)$. With half credit for ties, the reference-induced AUC error satisfies*

$$\mathbb{E}_R \left| \mathrm{AUC}_P(\widehat{P}_n f) - \mathrm{AUC}_P(P_Q f) \right| \le \mathbb{E}_{+,-} \exp\left(-\frac{n\Delta_f^2}{32B^2} \right). \tag{8}$$

If $\Pr(|\Delta_f| \le t) \le Ct^\alpha$ for $t > 0$ and $\alpha > 0$, the bound is at most

$$\inf_{t>0} \left\{ Ct^\alpha + \exp\left(-\frac{nt^2}{32B^2} \right) \right\}.$$

In particular, for $n \ge 2$ it is at most $C(32B^2 \log n/n)^{\alpha/2} + n^{-1}$.

Proof. For each fixed positive-negative pair, $W_f \in [-4B, 4B]$. If $\Delta_f > 0$, a change in its ordering requires $\widehat{\Delta}_f \le 0$. Hoeffding's one-sided bound gives probability at most $\exp[-n\Delta_f^2/(32B^2)]$ (Hoeffding, 1963). The same argument applies to $\Delta_f < 0$. At $\Delta_f = 0$, an absolute kernel difference is at most one, which preserves the bound. Integrating the absolute pairwise kernel difference proves the first inequality. Splitting into $|\Delta_f| \le t$ and its complement proves the second. The stated rate follows by substituting $t = (32B^2 \log n/n)^{1/2}$. □

Unbiased estimated margins can therefore yield biased estimated AUCs. Pairs close to a projected tie drive reference sensitivity. The margin condition excludes positive-negative tie mass. The concentration bound applies to predictors satisfying the stated boundedness condition. Empirical analyses use paired resampling of the fitted predictions. The finite-reference bias of $\widehat{D}_p(f)$ is bounded by the right side of (8). The corresponding bounds for $\widehat{G}_p$ and $\widehat{E}_p$ are the sum of the head and baseline bounds. These statements compare independent-sample expectations with the population-reference contrasts. Their scope is reference-estimation error under the stated fitting and sampling assumptions.

Proposition 4 (Paired AUC covariance)**.** *Condition on fitted models and reference R. The scoring sample is independent of both, with observations independent within and between its positive and negative class samples. Let $s = (f_h, f_c, \widehat{P}_n f_h, \widehat{P}_n f_c)$ and define the vector kernel*

$$K_k(x^+, x^-) = \mathbf{1}\{s_k(x^+) > s_k(x^-)\} + \tfrac{1}{2}\mathbf{1}\{s_k(x^+) = s_k(x^-)\}.$$

Suppose $n_+ \ge 1$ independent positive observations and $n_- \ge 1$ independent negative observations are scored. Let $\theta = \mathbb{E}K$ and $\widehat{\theta} = (n_+ n_-)^{-1} \sum_{i,j} K(X_i^+, X_j^-)$. Define

$$a(x^+) = \mathbb{E}_-[K(x^+, X^-)] - \theta, \quad b(x^-) = \mathbb{E}_+[K(X^+, x^-)] - \theta, \quad r = K - \theta - a - b.$$

Then $\mathbb{E}[\widehat{\theta} \mid R] = \theta(R)$ and

$$\begin{aligned} \mathrm{Cov}(\widehat{\theta} \mid R) &= \frac{\Sigma_a}{n_+} + \frac{\Sigma_b}{n_-} + \frac{\Sigma_r}{n_+ n_-}, \\ \Sigma_a = \mathbb{E}[aa^\top], \quad \Sigma_b &= \mathbb{E}[bb^\top], \quad \Sigma_r = \mathbb{E}[rr^\top]. \end{aligned} \tag{9}$$

Set

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$

The contrast vector $(G, G_p, E_p, D_p(f_h), D_p(f_c))^\top$ equals $L\theta$. Its estimator has conditional covariance

$$L\,\mathrm{Cov}(\widehat{\theta} \mid R)L^\top.$$

Proof. Conditional on R, the score functions are fixed and the two class samples are independent. Every kernel term has expectation θ. The two-sample Hoeffding decomposition gives

$$\widehat{\theta} - \theta = \frac{1}{n_+} \sum_i a(X_i^+) + \frac{1}{n_-} \sum_j b(X_j^-) + \frac{1}{n_+ n_-} \sum_{i,j} r(X_i^+, X_j^-).$$

The three terms are pairwise orthogonal in covariance. Residual terms sharing one observation also have zero covariance, by conditional centring. Summing the remaining diagonal covariance terms gives (9). Linear transformation gives the contrast covariance. □

This result applies established U-statistic theory and correlated-AUC inference (Hoeffding, 1948; DeLong et al., 1988) to the four shared score vectors. It yields the useful relation

$$\mathrm{Var}(\widehat{E}_p \mid R) = \mathrm{Var}(\widehat{G} \mid R) + \mathrm{Var}(\widehat{G}_p \mid R) - 2\,\mathrm{Cov}(\widehat{G}, \widehat{G}_p \mid R).$$

Standard DeLong covariance estimators target the leading terms asymptotically. Pointwise Wald or bootstrap intervals require their usual regularity conditions and have approximate coverage. For a random independent reference, total covariance also contains $\mathrm{Cov}_R\{\theta(R)\}$ through the law of total covariance. Resampling the scoring sample with projected predictions fixed estimates only scoring uncertainty conditional on that reference. Joint reference-and-scoring resampling recomputes every projected score and preserves cross-model pairing. If scoring observations also construct the reference, the independence assumption above fails. That design needs its own dependence analysis or a separate-reference evaluation.

For macro-AUC, the class-specific argument applies to each label. Cross-label covariance must be retained when those labelwise estimates share observations. An entity-level joint bootstrap preserves the pairing across labels, models and fitted seeds. Pooling all fold scores after cross-fitting creates comparisons between different random reference functions. An implementation should instead define its fold-specific AUC estimand and aggregation before scoring.

Interaction information and score sensitivity. For discrete variables, choose the co-information convention $I(U;V;Y) = I(U;V) - I(U;V \mid Y)$ (McGill, 1954). This quantity depends on the data law, whereas D_p also depends on the fitted function and score metric. For independent Rademacher U, V and $Y = \mathbf{1}\{UV = 1\}$, the co-information is -1 bit. The predictor $f = UV$ has $D_p = 1/2$, while a constant predictor has $D_p = 0$ under the same law. For $Y = \mathbf{1}\{U = 1\}$ and $f = U + \epsilon UV$, with $0 < \epsilon < 1$, co-information and D_p are both zero. Its squared projection residual is nevertheless ϵ^2, and its two-player SHAP interaction is $\epsilon UV/2$. Positive score scaling preserves both AUCs while rescaling the residual norm. The three diagnostics therefore quantify distinct properties of the task and predictor.

General scores and regression. For any specified scalar utility S, algebra gives

$$\begin{aligned} D_p(f) &= S(f) - S(Pf), & G &= S(f_h) - S(f_c), \\ G_p &= S(Pf_h) - S(Pf_c), & E_p &= D_p(f_h) - D_p(f_c), \\ G &= G_p + E_p. \end{aligned}$$

For regression, take $S(s) = -\mathbb{E}|Y - s(X)|$, so larger values retain the favourable direction. Conditional on fitted models and a fixed reference, define

$$T(X,Y) = \big(-|Y - f_h(X)|, -|Y - f_c(X)|, -|Y - Pf_h(X)|, -|Y - Pf_c(X)|\big)^\top.$$

For m independent scoring observations with finite second moments, $\widehat{\theta} = m^{-1} \sum_i T_i$ satisfies $\mathbb{E}\widehat{\theta} = \mathbb{E}T$ and $\mathrm{Cov}(L\widehat{\theta}) = L\,\mathrm{Cov}(T)L^\top/m$. These conditional statements allow the fixed reference to come from the training sample. They require the scoring sample to remain independent of training, selection and reference construction. For clustered video segments, independence applies at the source-video level and the sampling procedure must retain that clustering. The sample covariance divided by m estimates the individual-observation version without ignoring cross-model dependence. Correlation, macro-F1 and other nonlinear metrics require metric-specific inference or a suitable paired resampling procedure. The score identity alone supplies no uncertainty guarantee.

The projection-induced MAE change satisfies

$$|D_p(f)| \le \mathbb{E}_P|f(X) - Pf(X)| \le \{\mathbb{E}_P[(f(X) - Pf(X))^2]\}^{1/2},$$

by the reverse triangle and Cauchy–Schwarz inequalities. This uses the evaluation law P, while the ANOVA norm in Proposition 1 uses Q. If the feature law P_X has density bounded by M relative to Q, then $|D_p(f)| \le \sqrt{M}\|f - P_Q f\|_Q$ for the population projection. The density condition is substantive when cross-modal dependence concentrates the evaluation law.

Protocol implication. Reserve independent reference and scoring samples before inspecting task outcomes. Use the same reference rows, scoring rows, partitions and metric for every selected model. Report G, G_p, E_p and both D_p values with their paired uncertainty. State whether uncertainty conditions on the reference or includes its re-estimation. Report near-tie mass and reference-size sensitivity alongside the principal contrasts.

B Decision targets and partition design

Interpreting a score sensitivity. The quantities D_p and E_p describe changes in a specified task utility under a specified projection. A positive $D_p(f)$ means that the original predictor scores higher than its additive projection on the evaluation distribution. The residual's functional magnitude is measured separately through an appropriate norm or interaction attribution. For example, multiplying scalar logits by a positive constant preserves AUC while changing the residual norm. Changing the label mechanism can alter D_p while preserving the prediction function and its residual. Consequently, utility, function structure and deployment performance require their corresponding evaluation targets.

Choosing an evaluation objective. For ordinary deployment, model selection uses validation utility and the deployment resource budget. If an application explicitly values both original and projected performance, it can declare

$$U_\lambda(f) = (1 - \lambda)S(f) + \lambda S(P_p^R f), \qquad 0 \le \lambda \le 1. \tag{10}$$

This objective averages utilities obtained with two evaluation procedures. The coefficient λ specifies the value assigned to projected performance and is fixed before model selection. Both candidate models must first satisfy the same declared deployment constraints. Under a common baseline, score scale, partition and reference, its relative value is

$$U_\lambda(f_h) - U_\lambda(f_c) = G - \lambda E_p.$$

Consider two illustrative heads with $(G, G_p, E_p) = (2.0, 1.8, 0.2)$ and $(2.0, 0.1, 1.9)$, with all gaps in percentage points. Their relative utilities are $2.0 - 0.2\lambda$ and $2.0 - 1.9\lambda$. The heads tie on these point estimates at $\lambda = 0$, and the first has a 1.7λ advantage for $\lambda > 0$. For an interaction-focused analysis, the second head displays greater excess task sensitivity under the specified reference. Final comparisons use an independent scoring set after validation selection. Applying the same $G - \lambda E_p$ contrast within each paired scoring resample gives an interval for the combined utility advantage. Deployment with missing modalities or distribution shift requires direct evaluation under those conditions. For nonlinear metrics, evaluating an average of predictions also defines a different utility.

Interpreting the GMU comparison. For matched VQA concatenation, $E_p < 0$ means that the baseline loses more task utility under the shared projection. The positive G_p records the resulting advantage between the two projected predictors. A comparison of functional additivity instead examines both residuals under the same specified reference law.

Selecting partitions for many blocks. For $k \ge 2$ input blocks, there are $2^{k-1} - 1$ unordered, nonempty bipartitions. A practical initial set uses each block versus all remaining blocks, with complementary duplicates removed. Domain-defined groups and subtrees of a specified fusion hierarchy can supply additional cuts, each against all remaining blocks. Selecting a linear number of these cuts from domain knowledge before scoring preserves an $O(k)$ partition budget. All selected cuts use common reference rows and shared scoring resamples. Their contrasts remain separate because one functional interaction can cross several cuts. Under a common fully factorised reference law, each residual contains all ANOVA terms spanning its cut. When groupwise references preserve dependencies within either side, each cut has its own product reference law.

Choosing reference size and composition. The reference law identifies which cross-paired predictions contribute to the diagnostic. Sampling observations uniformly gives frequent source groups greater weight than sampling groups uniformly. The appropriate weighting follows the application population and is declared before comparing task scores. For numerical precision, a validation-only pilot can double reference size across several draws and compare changes in G_p and E_p. The pilot uses a predetermined tolerance in task-utility units and a specified maximum reference budget. The chosen size, law and remaining draw variation accompany the final report. The empirical tolerance describes the pilot's stability, while formal guarantees require the assumptions of the corresponding reference analysis.

C Reusable diagnostic procedure and reporting

Algorithm: Shared-reference projection diagnosis

Input: Fixed predictors f_h, f_c, labelled evaluation units D, partition $p = U|V$, reference bank R, metric S, and optional cluster IDs.
Before evaluation: Fix selection rules, capacity comparisons, score scale, reference construction, and the comparison family.

1. Construct a shared reference $Q = \widehat{Q}_U \otimes \widehat{Q}_V$ from R. For sampled integration, draw marginal indices independently and fix them.
2. For each $f \in \{f_h, f_c\}$, evaluate $f(x_i)$ and
$$\widehat{P}_p f(x_i) = \widehat{\mathbb{E}}_v f(x_{i,U}, v) + \widehat{\mathbb{E}}_u f(u, x_{i,V}) - \widehat{\mathbb{E}}_{u,v} f(u, v).$$
Use the same integration mode and reference draws for both predictors.
3. Compute the four task scores and $G = S(f_h) - S(f_c)$, $G_p = S(\widehat{P}_p f_h) - S(\widehat{P}_p f_c)$, and $E_p = G - G_p$. Also retain both own-model sensitivities.
4. Draw evaluation rows or clusters with replacement. Apply each draw to all four fixed score arrays and every fitted seed. Compute within-seed contrasts and then their seed mean.
5. Return raw and projected scores, paired contrasts, pointwise intervals, and the task, selection, capacity, reference, and uncertainty record.

Figure 4: A reusable comparison based on EMAP (Hessel & Lee, 2020). Cluster resampling retains every row of each sampled cluster. Test resampling holds fitted models, selection, and the reference fixed.

The library implements the shared-reference score comparison with array-based predictors, user-defined metrics and optional cluster identifiers. Exact reference integration and paired seed aggregation share the manuscript's conventions. The reporting record includes task and split identifiers, score scale, model and baseline definitions, trainable counts, selection budgets, reference identifiers, all four scores, all contrasts and uncertainty conditioning. The release supplies an installable package, command-line entry point and worked examples. From saved VQA predictions, the released library reproduces all four paired comparisons and their interval endpoints within 3.4×10^{-16} in accuracy units. It also reproduces 360 regression contrast means and interval endpoints across MOSI and MOSEI within 7.0×10^{-16} MAE units. Tests check the score identity, paired resampling, projection geometry, reference handling and metric agreement.

D Capacity control protocol and complete results

The capacity study retains the original movie banks, training splits and fitted Mamba-style checkpoints. For concatenation with input dimension D, output dimension 23 and hidden width w, the parameter count is $w(D + 24) + 23$. The nearest integer width minimises its difference from the corresponding Mamba-style count. Exact ties select the smaller width. The resulting counts are 2,307,959 versus 2,308,119 on w2v and 2,427,323 versus 2,427,927 on ST.

The four widths are 256, 512, 1,024 and the matched width for each bank. Each width uses learning rates $\{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$ and seeds 0, 1 and 2. Every new model follows the 30-epoch AdamW recipe with batch size 256 and weight decay 10^{-4}. The input and hidden dropout rates remain 0.3 and 0.2. The best validation macro-AUC selects the earliest tied epoch. The largest three-seed mean validation AUC then selects the learning rate at each width. An exact configuration tie selects the smaller learning rate. All eight width selections chose 3×10^{-4}. Thus the width-256 point in this curve uses its own validation-selected recipe. Default concatenation uses learning rate 10^{-3}.

Thirty-six width-256/512 fits were reused after checking their data and checkpoint hashes. Thirty-six width-1,024/matched fits were trained anew. Held-out evaluation uses the eight configurations recorded in Section 4. Every selected model was evaluated with the same seed ordering. Mamba-style retains its earlier fixed training recipe. This controls parameter number under the stated optimisation and selection procedures.

For the matched comparison, each model uses the complete $7,799^2$ test cross-product at three partitions. Cached input transformations were checked against direct predictions, with maximum difference 3.82×10^{-6}. Scores remain unrounded, and grid means accumulate in float64. The 1,000 paired row resamples use seed 20260912 across all models, fitted seeds and partitions within each bank. All replicates retain both classes for every genre. These resampling intervals hold the observed full-test reference fixed. The independent scoring-sample results in Section A concern a separate reference design.

Table 5: Capacity curve with validation-selected learning rate at every width. Macro-AUC is the mean $\pm$ sample SD of three seeds. All selected learning rates were 3×10^{-4}. Matched widths minimise the actual parameter difference from Mamba-style.

Text	Width	Trainable parameters	Macro-AUC
w2v	256	356,375	0.8562 ± 0.0003
w2v	512	712,727	0.8567 ± 0.0006
w2v	1024	1,425,431	0.8557 ± 0.0004
w2v	1658†	2,307,959	0.8559 ± 0.0004
ST	256	476,183	0.9328 ± 0.0005
ST	512	952,343	0.9332 ± 0.0002
ST	1024	1,904,663	0.9332 ± 0.0003
ST	1305†	2,427,323	0.9335 ± 0.0005

† Parameter-matched width.

Table 6: Own-model projection sensitivities for the capacity comparison. Entries are three-seed means and pointwise paired resampling intervals in AUC percentage points. The fitted models and full-test reference remain fixed. Intervals use capacity-study bootstrap seed 20260912. They differ slightly from Table 47, which uses fixed-checkpoint seed 20260910.

Text	Partition	D_p(Mamba-style)	D_p(matched concat)
w2v	$ab \mid c$	0.289 [0.208, 0.376]	0.197 [0.168, 0.225]
	$ac \mid b$	0.328 [0.254, 0.411]	0.288 [0.227, 0.347]
	$a \mid bc$	0.115 [0.033, 0.211]	0.144 [0.088, 0.202]
ST	$ab \mid c$	0.181 [0.138, 0.226]	0.070 [0.058, 0.083]
	$ac \mid b$	0.156 [0.113, 0.201]	0.134 [0.096, 0.171]
	$a \mid bc$	0.230 [0.178, 0.284]	0.172 [0.132, 0.209]

E CAPACITY CONTROL WITH A TRAINING REFERENCE

This secondary analysis followed the completed full-test-reference capacity comparison. A dated amendment specifies the reference construction and inference procedure. The analysis uses the capacity-study Mamba-style and matched concatenation checkpoints for both banks and all three seeds. For each bank, 256 training movies were sampled without replacement using seed 8101. Their complete empirical product defined the reference for all models, seeds and partitions. The scoring sample comprised the original 7,799 test movies.

Projection sums used float64 accumulation, including the complete reference-by-reference grand mean. Direct cross-bank predictions verified the cached computation. Recomputed original logits agreed exactly with the stored original predictions. The analysis used 1,000 paired movie-bootstrap draws with seed 20260912. Each draw was shared across models, fitted seeds and partitions. Metrics were computed within seed before averaging. The intervals condition on these models and the fixed training reference. Both banks retained all 1,000 draws.

The original macro-AUCs remained 0.8710 ± 0.0008 for Mamba-style and 0.8559 ± 0.0004 for matched concatenation on w2v. The corresponding ST scores were 0.9344 ± 0.0007 and 0.9335 ± 0.0005. The original gaps and their intervals are unchanged from the primary capacity comparison.

Table 7: Secondary capacity comparison under the 256-training-movie reference. Projected macro-AUCs are mean ± sample SD across three seeds. Gaps are AUC percentage points with pointwise 95% paired percentile intervals. The fitted models and reference remain fixed during resampling.

Bank	Partition	$S(P_p f_h)$	$S(P_p f_c)$	G_p	E_p
w2v	$ab \mid c$	0.8680 ±0.0002	0.8539 ±0.0004	1.413 [1.146, 1.660]	0.091 [0.021, 0.169]
w2v	$ac \mid b$	0.8676 ±0.0002	0.8531 ±0.0004	1.451 [1.180, 1.707]	0.053 [−0.026, 0.139]
w2v	$a \mid bc$	0.8698 ±0.0009	0.8545 ±0.0004	1.524 [1.264, 1.766]	−0.019 [−0.108, 0.075]
ST	$ab \mid c$	0.9325 ±0.0007	0.9328 ±0.0004	−0.030 [−0.148, 0.096]	0.116 [0.076, 0.157]
ST	$ac \mid b$	0.9327 ±0.0005	0.9321 ±0.0004	0.060 [−0.060, 0.181]	0.026 [−0.015, 0.067]
ST	$a \mid bc$	0.9320 ±0.0005	0.9318 ±0.0004	0.028 [−0.078, 0.148]	0.058 [0.006, 0.106]

Table 8: Own-model projection sensitivities under the same training reference. Values are AUC percentage points with pointwise 95% paired percentile intervals.

Bank	Partition	$D_p(f_h)$	$D_p(f_c)$
w2v	$ab \mid c$	0.292 [0.212, 0.382]	0.201 [0.173, 0.230]
w2v	$ac \mid b$	0.331 [0.255, 0.413]	0.278 [0.216, 0.337]
w2v	$a \mid bc$	0.119 [0.037, 0.213]	0.138 [0.082, 0.194]
ST	$ab \mid c$	0.186 [0.142, 0.230]	0.070 [0.058, 0.083]
ST	$ac \mid b$	0.162 [0.117, 0.207]	0.135 [0.097, 0.172]
ST	$a \mid bc$	0.232 [0.179, 0.287]	0.174 [0.134, 0.212]

The release records this analysis in the dated training-reference capacity amendment. Its aggregate output preserves all six partition-matched comparisons alongside the primary analysis.

F AUDIOVISUAL-LANGUAGE SENTIMENT PROTOCOL

Source and representations. The experiments use the MMSA aligned-50 feature releases and model code at commit `a94e65d07fa1ae0d44e552390074b29b0898edfd` (THUIAR, 2025). File SHA-256 digests match the author's published feature manifest. The author-provided splits have disjoint clip and source-video identifiers. Both tasks retain 50 text steps and average

all 50 audio and visual steps before their corresponding subnetworks. Nonfinite feature values are mapped to zero, and all rows are retained. The text representations are frozen input features. LSTM and other unimodal subnetworks train jointly with each fusion readout. The aligned release, preprocessing and training configuration define the evaluated implementation. The public code includes the original feature-file hashes, model sources and full configuration inventory.

Predefined models and capacity. LMF uses its modality-specific low-rank tensor factors (Liu et al., 2018). The ranks are 3 for MOSI and 4 for MOSEI. TFN computes an augmented outer product and nonlinear regression readout (Zadeh et al., 2017). For each model, its matched concatenation copies the same initial unimodal subnetworks and replaces the fusion readout with a two-hidden-layer ReLU MLP. The models subsequently train independently. All matched parameters participate in prediction. LMF matched concatenation uses hidden widths 3/2 on MOSI and 4/2 on MOSEI. TFN matched concatenation uses widths 2,265/3,775 and 1,349/3,078, respectively. Both hidden widths are searched over integers from 1 to 8,192. The selection minimises the readout parameter difference, then favours more balanced widths, a smaller width sum and the smaller first width. The LMF output is linear. Both TFN and its matched concatenation retain the same $6\sigma(z) - 3$ output transformation. TFN's two fixed range constants are excluded from trainable counts. Shared concatenation uses temporally averaged input features, a 256/128 ReLU MLP and a linear output. It supplies a common comparison across the architecture-specific matched concatenation models.

Training settings. All fits use one Adam parameter group, maximum 60 epochs, MAE loss and global gradient-norm clipping at 5. Validation MAE selects the first strictly best checkpoint. Training stops after eight epochs without improvement. The last singleton batch is merged into the preceding batch so each training observation appears once per epoch. Each fitted seed has a separate deterministic batch-order generator. Models use float32, highest matrix-multiplication precision, TF32 disabled and deterministic algorithms. The original source learning rates, per-family batch sizes and regularisation settings are listed below. Each matched concatenation model inherits its corresponding family's batch size, weight decay and dropout settings.

Table 9: Sentiment training settings. The source learning rate and two smaller rates form each model's three-candidate search. Dropout entries follow the audio, visual, text and fusion order in the MMSA configuration.

Dataset	Family	Source LR	Batch	Weight decay	Dropout
MOSI	LMF / matched	0.001	64	0.005	0.3/0.3/0.3/0.3
MOSI	TFN / matched	0.001	32	0	0.3/0.3/0.3/0.5
MOSEI	LMF / matched	0.001	32	0.001	0.2/0.2/0.2/0.2
MOSEI	TFN / matched	0.005	128	0	0.4/0.4/0.4/0.4
Both	Shared concat	0.001	64	0.001	0.2

Protocol sequence and learning-rate selection. The initial protocol specified 30 source-rate fits across two datasets, five models and three seeds before fitting began. After examining validation optimisation, the pre-test amendment described in Section 4 added candidates 10^{-4} and 3×10^{-4} alongside each model's source learning rate. This added 60 fits with all other training rules unchanged. For each of ten dataset/model combinations, mean validation MAE across the three seeds selected one learning rate, with exact ties favouring the smaller rate. The selected configurations from Section 4 were used for held-out evaluation. The original source-rate analysis is reported separately.

Projection and uncertainty. For each dataset, 256 training rows are sampled without replacement using seed 8101 and sorted by their source row index. All models and fitted seeds use the same rows for $T|AV$, $A|TV$ and $V|TA$. Each grouped pair retains its within-segment pairing. For every scored segment, projection evaluates both reference cross terms and the full 256×256 grand mean. Forward passes use float32 and means accumulate in float64. Both diagonal and crossed-input computations are checked against the original fitted model. Regression projection retains the grand mean because an output shift can change MAE.

The utility is negative MAE, so positive G and G_p favour the fusion head. Two thousand bootstrap replicates sample original videos with replacement using seed 8102. Each selected video contributes all its segments, and the resampled loss is divided by the resulting segment count. Each

draw is shared across models, partitions and seeds. Seed-specific metric differences are averaged before percentile calculation. Intervals condition on the fitted models, selected configurations and fixed training reference. The primary comparisons use each head's matched concatenation. Shared concatenation defines secondary comparisons. Secondary metrics include Pearson correlation, binary accuracy and weighted F1 excluding zero targets, and seven-class accuracy after clipping and rounding to $[-3, 3]$.

F.1 Complete sentiment results

LMF matched and TFN matched denote their respective parameter-matched concatenation controls. Shared concat denotes shared concatenation on temporally averaged features.

Table 10: Validation-selected learning-rate results for the complete model roster. MAE is the mean $\pm$ sample standard deviation across three seeds. Fixed TFN output-range constants are excluded from trainable counts.

Dataset	Model	Trainable parameters	MAE ↓
MOSI	LMF	505,071	1.126 ± 0.290
MOSI	LMF matched	505,072	1.292 ± 0.309
MOSI	TFN	9,497,043	0.944 ± 0.010
MOSI	TFN matched	9,497,041	0.978 ± 0.040
MOSI	Shared concat	236,289	0.949 ± 0.024
MOSEI	LMF	508,475	0.565 ± 0.005
MOSEI	LMF matched	508,475	0.665 ± 0.150
MOSEI	TFN	5,042,539	0.578 ± 0.009
MOSEI	TFN matched	5,042,538	0.578 ± 0.003
MOSEI	Shared concat	257,793	0.576 ± 0.003

Table 11: Validation-only learning-rate selection. All three MAE columns report arithmetic mean validation MAE across seeds 0, 1 and 2. Complete per-seed validation values are supplied in the sealed selection manifest.

Dataset	Model	Source LR	Source	10^{-4}	3×10^{-4}	Chosen LR
MOSI	LMF	0.001	1.0700	1.1033	1.0749	0.001
MOSI	LMF matched	0.001	1.2490	1.3343	1.3085	0.001
MOSI	TFN	0.001	0.8652	0.9220	0.8992	0.001
MOSI	TFN matched	0.001	0.8803	0.9376	0.8830	0.001
MOSI	Shared concat	0.001	0.9153	0.9228	0.9215	0.001
MOSEI	LMF	0.001	0.5397	0.5529	0.5483	0.001
MOSEI	LMF matched	0.001	0.6280	0.6363	0.6289	0.001
MOSEI	TFN	0.005	0.5516	0.5628	0.5589	0.005
MOSEI	TFN matched	0.005	1.3050	0.5641	0.5548	0.0003
MOSEI	Shared concat	0.001	0.5459	0.5488	0.5477	0.001

Table 12: Validation-selected learning rates. Secondary comparisons against shared concatenation. Estimates and pointwise 95% intervals are in 10^{-3} MAE units, rounded to two decimal places.

Dataset	Model	Partition	G	G_p	E_p
			-176.84	-172.53	-4.30
MOSI	LMF	$T \mid AV$	$[-223.87, -129.15]$	$[-219.74, -124.94]$	$[-11.84, 2.74]$
			-176.84	-176.72	-0.12
		$A \mid TV$	$[-223.87, -129.15]$	$[-224.59, -129.21]$	$[-2.72, 2.46]$
			-176.84	-172.31	-4.53
		$V \mid TA$	$[-223.87, -129.15]$	$[-218.87, -123.95]$	$[-11.04, 1.53]$
			5.36	12.39	-7.03
	TFN	$T \mid AV$	$[-33.78, 40.97]$	$[-22.85, 46.92]$	$[-17.04, 2.38]$
			5.36	3.20	2.16
		$A \mid TV$	$[-33.78, 40.97]$	$[-34.81, 38.04]$	$[-2.95, 6.80]$
			5.36	15.34	-9.98
		$V \mid TA$	$[-33.78, 40.97]$	$[-22.19, 51.06]$	$[-18.81, -1.47]$
			11.09	15.20	-4.11
MOSEI	LMF	$T \mid AV$	$[3.12, 19.25]$	$[6.69, 23.93]$	$[-8.15, -0.30]$
			11.09	14.84	-3.76
		$A \mid TV$	$[3.12, 19.25]$	$[6.73, 23.39]$	$[-6.26, -1.24]$
			11.09	13.51	-2.43
		$V \mid TA$	$[3.12, 19.25]$	$[5.18, 21.82]$	$[-5.52, 0.40]$
			-2.27	-1.29	-0.98
	TFN	$T \mid AV$	$[-10.21, 5.85]$	$[-9.62, 6.81]$	$[-4.79, 2.63]$
			-2.27	0.38	-2.65
		$A \mid TV$	$[-10.21, 5.85]$	$[-8.14, 8.86]$	$[-5.22, -0.21]$
			-2.27	-2.56	0.29
		$V \mid TA$	$[-10.21, 5.85]$	$[-10.54, 5.27]$	$[-2.52, 2.94]$

Table 13: Validation-selected learning rates. Model-specific D_p estimates and pointwise 95% video-bootstrap intervals in 10^{-3} MAE units. Values are rounded to two decimal places; scientific notation retains smaller nonzero values and interval endpoints. Positive values indicate increased MAE after projection.

Dataset	Model	$D_{T\|AV}$	$D_{A\|TV}$	$D_{V\|TA}$
		-4.53	-0.31	-4.65
MOSI	LMF	$[-12.36, 2.29]$	$[-2.36, 1.62]$	$[-11.55, 1.46]$
		-0.72	2.26×10^{-8}	-0.72
MOSI	LMF matched	$[-1.88, 0.31]$	$[-5.08, 4.54] \times 10^{-6}$	$[-1.88, 0.31]$
		-7.26	1.97	-10.10
MOSI	TFN	$[-17.20, 1.56]$	$[-1.77, 5.60]$	$[-19.49, -1.32]$
		-0.72	0.27	-1.06
MOSI	TFN matched	$[-4.28, 2.40]$	$[-0.88, 1.44]$	$[-4.38, 1.92]$
		-0.23	-0.19	-0.11
MOSI	Shared concat	$[-4.57, 4.05]$	$[-2.47, 2.30]$	$[-2.91, 2.60]$
		-0.09	-0.28	0.06
MOSEI	LMF	$[-2.22, 1.98]$	$[-0.93, 0.39]$	$[-1.87, 1.97]$
		-9.92×10^{-7}	-9.92×10^{-7}	-9.92×10^{-7}
MOSEI	LMF matched	$[-2.13, 0.14] \times 10^{-6}$	$[-2.13, 0.14] \times 10^{-6}$	$[-2.13, 0.14] \times 10^{-6}$
		3.04	0.82	2.78
MOSEI	TFN	$[0.99, 5.15]$	$[-0.58, 2.16]$	$[1.42, 4.12]$
		3.38	1.09	2.72
MOSEI	TFN matched	$[1.34, 5.46]$	$[0.64, 1.53]$	$[0.85, 4.62]$
		4.02	3.47	2.49
MOSEI	Shared concat	$[0.22, 7.98]$	$[0.84, 6.04]$	$[-0.43, 5.69]$

Table 14: Validation-selected learning rates. Secondary metrics are mean ± sample standard deviation across three seeds. Acc2 and weighted F1 exclude zero targets. Acc7 rounds clipped predictions and targets to $[-3, 3]$. $\mathrm{n/a}^{\dagger}$ marks a seed group containing at least one constant-prediction run. Every seed contributes to MAE, accuracy and F1. Accuracy and F1 use percentages.

Dataset	Model	Corr.	Acc2 (%)	F1 (%)	Acc7 (%)
MOSI	LMF	$\mathrm{n/a}^{\dagger}$	65.701 ± 20.358	60.059 ± 30.316	28.377 ± 11.209
MOSI	LMF matched	$\mathrm{n/a}^{\dagger}$	54.268 ± 20.859	42.814 ± 30.728	22.643 ± 12.456
MOSI	TFN	0.667 ± 0.014	79.268 ± 1.000	79.257 ± 0.998	35.569 ± 1.749
MOSI	TFN matched	0.665 ± 0.020	77.287 ± 1.904	77.410 ± 1.887	34.257 ± 3.798
MOSI	Shared concat	0.675 ± 0.016	77.998 ± 0.616	78.111 ± 0.595	33.090 ± 2.723
MOSEI	LMF	0.731 ± 0.005	83.480 ± 0.374	83.435 ± 0.282	52.136 ± 0.355
MOSEI	LMF matched	$\mathrm{n/a}^{\dagger}$	76.050 ± 11.470	70.899 ± 19.421	47.922 ± 5.715
MOSEI	TFN	0.714 ± 0.005	80.692 ± 2.734	80.448 ± 2.216	51.248 ± 0.533
MOSEI	TFN matched	0.710 ± 0.003	82.315 ± 1.066	82.201 ± 0.886	50.891 ± 0.521
MOSEI	Shared concat	0.716 ± 0.002	83.269 ± 0.361	83.195 ± 0.420	51.928 ± 0.511

Source-rate comparison. Validation selection retains the source learning rate for nine of ten model/dataset definitions. Their source-rate and selected predictions coincide. The following tables report MOSEI TFN and its matched concatenation at the source rate of 0.005. The shared-baseline results in Table 12 apply to both analyses.

Table 15: MOSEI TFN and its matched concatenation at their original learning rate. MAE is the mean ± sample standard deviation across three seeds. Fixed TFN output-range constants are excluded from trainable counts.

Dataset	Model	Trainable parameters	MAE ↓
MOSEI	TFN	5,042,539	0.578 ± 0.009
MOSEI	TFN matched	5,042,538	1.331 ± 1.322

Table 16: MOSEI TFN and its matched concatenation at their original learning rate. Fusion models versus their matched concatenation. Estimates and pointwise 95% video-bootstrap intervals are in MAE units.

Dataset	Model	Partition	G	G_p	E_p
MOSEI	TFN	$T \mid AV$	0.753 $[0.723, 0.780]$	0.752 $[0.722, 0.779]$	0.001 $[-0.001, 0.003]$
MOSEI	TFN	$A \mid TV$	0.753 $[0.723, 0.780]$	0.753 $[0.723, 0.780]$	-0.00027 $[-0.002, 0.001]$
MOSEI	TFN	$V \mid TA$	0.753 $[0.723, 0.780]$	0.751 $[0.722, 0.779]$	0.001 $[0.0001, 0.002]$

Table 17: MOSEI TFN and its matched concatenation at their original learning rate. Model-specific D_p estimates and pointwise 95% video-bootstrap intervals in MAE units. Positive values indicate increased MAE after projection.

Dataset	Model	$D_{T\|AV}$	$D_{A\|TV}$	$D_{V\|TA}$
MOSEI	TFN	0.003 $[0.001, 0.005]$	0.001 $[-0.001, 0.002]$	0.003 $[0.001, 0.004]$
MOSEI	TFN matched	0.002 $[0.001, 0.003]$	0.001 $[0.001, 0.001]$	0.002 $[0.001, 0.003]$

Table 18: MOSEI TFN and its matched concatenation at their original learning rate. Secondary metrics are mean ± sample standard deviation across three seeds. Acc2 and weighted F1 exclude zero targets. Acc7 rounds clipped predictions and targets to $[-3, 3]$. $\mathrm{n/a}^{\dagger}$ marks a seed group containing at least one constant-prediction run. Every seed contributes to MAE, accuracy and F1. Accuracy and F1 use percentages.

Dataset	Model	Corr.	Acc2 (%)	F1 (%)	Acc7 (%)
MOSEI	TFN	0.714 ± 0.005	80.692 ± 2.734	80.448 ± 2.216	51.248 ± 0.533
MOSEI	TFN matched	$\mathrm{n/a}^{\dagger}$	74.867 ± 10.480	69.967 ± 18.594	34.607 ± 29.208

F.2 SEED-SPECIFIC HELD-OUT METRICS

The tables retain every predefined seed. Pearson correlation is reported for seeds with nonzero prediction variance. MOSI LMF seed 0 and its matched concatenation seeds 1 and 2 are constant in both analyses. MOSEI LMF-matched seed 0 is constant in both analyses. The original-rate MOSEI TFN-matched seed 1 is also constant. The validation-selected TFN-matched configuration has nonconstant predictions for every seed. Nine model/dataset configurations select their original rate and share the same seed-specific metrics in both analyses. The additional source-rate table records the one configuration that differs.

Table 19: MOSI. Validation-selected learning rates. Each row is one predefined seed. $\mathrm{n/a}^{\dagger}$ marks a seed with zero prediction variance. All other metrics retain these predictions. Acc2 and weighted F1 exclude zero targets.

Model	Seed	MAE	Corr.	Acc2 (%)	F1 (%)	Acc7 (%)
LMF	0	1.4579	$\mathrm{n/a}^{\dagger}$	42.23	25.07	15.45
LMF	1	0.9249	0.6725	78.51	78.59	35.42
LMF	2	0.9949	0.6411	76.37	76.51	34.26
LMF matched	0	0.9378	0.6444	78.35	78.30	37.03
LMF matched	1	1.5063	$\mathrm{n/a}^{\dagger}$	42.23	25.07	15.45
LMF matched	2	1.4307	$\mathrm{n/a}^{\dagger}$	42.23	25.07	15.45
TFN	0	0.9547	0.6710	80.18	80.19	35.57
TFN	1	0.9349	0.6784	79.42	79.37	33.82
TFN	2	0.9415	0.6513	78.20	78.21	37.32
TFN matched	0	0.9943	0.6421	75.15	75.29	36.73
TFN matched	1	0.9326	0.6733	77.90	78.01	36.15
TFN matched	2	1.0076	0.6791	78.81	78.92	29.88
Shared concat	0	0.9496	0.6819	78.35	78.49	32.65
Shared concat	1	0.9249	0.6864	78.35	78.42	36.01
Shared concat	2	0.9728	0.6570	77.29	77.42	30.61

Table 20: MOSEI. Validation-selected learning rates. Each row is one predefined seed. $\mathrm{n/a}^{\dagger}$ marks a seed with zero prediction variance. All other metrics retain these predictions. Acc2 and weighted F1 exclude zero targets.

Model	Seed	MAE	Corr.	Acc2 (%)	F1 (%)	Acc7 (%)
LMF	0	0.5616	0.7346	83.13	83.14	52.16
LMF	1	0.5628	0.7328	83.87	83.71	52.48
LMF	2	0.5704	0.7257	83.43	83.46	51.77
LMF matched	0	0.8384	$\mathrm{n/a}^{\dagger}$	62.85	48.51	41.36
LMF matched	1	0.5801	0.7137	83.60	83.25	50.59
LMF matched	2	0.5780	0.7092	81.70	80.94	51.81
TFN	0	0.5883	0.7097	77.63	78.00	50.63
TFN	1	0.5764	0.7120	81.56	81.02	51.53
TFN	2	0.5703	0.7195	82.88	82.32	51.58
TFN matched	0	0.5746	0.7135	83.24	83.00	50.74
TFN matched	1	0.5800	0.7094	81.15	81.25	51.47
TFN matched	2	0.5780	0.7068	82.55	82.35	50.46
Shared concat	0	0.5735	0.7190	83.60	83.67	51.84
Shared concat	1	0.5794	0.7154	82.88	82.89	51.47
Shared concat	2	0.5752	0.7143	83.32	83.03	52.48

Table 21: MOSEI. Original-rate TFN-matched concatenation. Each row is one predefined seed. $\mathrm{n/a}^{\dagger}$ marks a seed with zero prediction variance. All other metrics retain these predictions. Acc2 and weighted F1 exclude zero targets.

Model	Seed	MAE	Corr.	Acc2 (%)	F1 (%)	Acc7 (%)
TFN matched	0	0.5678	0.7210	82.11	81.42	51.47
TFN matched	1	2.8577	$\mathrm{n/a}^{\dagger}$	62.85	48.51	0.88
TFN matched	2	0.5670	0.7289	79.64	79.97	51.47

G BALANCED BINARY ABSTRACT SCENES VQA

The official 2017 file release supplies the balanced binary abstract-scene task introduced by Zhang et al. (2016). All 22,055 official training questions and 11,328 official validation questions are retained. They use 20,629 and 10,696 RGB images. Original source-scene identity joins questions with their complementary scenes. There are 10,295 complementary pairs in training and 5,328 in held-out scoring. A grouped 85%/15% split with seed 20260912 yields 18,745 training and 3,310 internal validation questions. Their 8,783 and 1,551 scene groups are disjoint. The held-out set contains 5,368 source-scene groups, all disjoint from official training. No answer or identifier enters either feature vector.

Features and training. ResNet-18 uses its frozen ImageNet1K V1 weights and 512-dimensional global average pooling (He et al., 2016). RGB preprocessing resizes the shorter side to 256 pixels, centre-crops 224 pixels and uses ImageNet normalisation. The shared frozen feature encoder has 11,176,512 parameters. Training-question means and standard deviations standardise the image features. TF-IDF uses unigrams and bigrams, minimum document frequency 2, at most 4,096 features, sublinear term frequency and L2 normalisation. Vocabulary and IDF use internal training only. Each of five model definitions receives learning rates 3×10^{-4} and 10^{-3} and seeds 0, 1 and 2. All 30 fits use 30 epochs, batch size 256, AdamW and weight decay 10^{-4}. The loss is mean binary cross-entropy over two logits ordered as no and yes, using the official VQA-credit target for each candidate answer. Validation VQA accuracy selects the earliest best epoch. The three-seed mean validation score selects the learning rate, with the smaller rate breaking exact ties. Official validation answers were acquired after configuration selection.

Head definitions. Concatenation uses input dropout 0.2, one hidden ReLU layer, hidden dropout 0.2 and two output logits. MLB-style applies two width-256 affine projections and tanh activations, then a Hadamard product and linear readout. GMU-style uses the same modality projections and a width-256 sigmoid gate computed from concatenated inputs. It combines the two projections through complementary gate weights. Both heads use input and readout dropout 0.2. They implement pooled fusion mechanisms on shared frozen features. The nearest real-parameter concat match has width 256 for MLB-style and 512 for GMU-style. Their relative parameter differences are 0.0217% and 0.0108%.

Projection and inference. The image-versus-question projection acts separately on both raw logits. Its reference contains 256 internal-training question rows sampled without replacement with seed 8101. It uses both evaluation-reference cross terms and the complete 256×256 reference grand mean, with float32 forward computation and float64 accumulation. Argmax predicts no or yes, with exact ties assigned to no. The score uses the official leave-one-annotator-out credit across ten answers. Agreement with the official evaluator was exact for 296 candidate-answer checks, including every possible binary vote count. The 1,000 bootstrap replicates sample source-scene clusters with replacement. Every selected cluster contributes all its original and complementary question rows. Weighted accuracy divides the weighted credit total by the total row weight. Each draw is shared across models and fitted seeds, and differences precede seed averaging. The percentile intervals condition on fixed fits, chosen configurations and empirical reference. They provide pointwise uncertainty estimates.

Table 22: Selected VQA models. Accuracy is expressed as a percentage, with mean ± sample SD across three seeds. Parameter counts cover the trainable head.

Model	Parameters	Learning rate	Original	Projected
Concat-128	590,210	0.001	64.89 ± 0.27	64.76 ± 0.24
MLB-style	1,180,674	0.0003	67.10 ± 0.14	64.57 ± 0.21
GMU-style	2,360,578	0.0003	65.34 ± 0.12	65.25 ± 0.24
Concat-256	1,180,418	0.001	64.80 ± 0.30	64.57 ± 0.22
Concat-512	2,360,834	0.001	64.89 ± 0.36	64.53 ± 0.56

Table 23: VQA original and projected accuracy for every selected seed, in percent.

Model	Seed	Original	Projected
Concat-128	0	65.126236	64.950565
Concat-128	1	64.597458	64.494174
Concat-128	2	64.960275	64.841102
MLB-style	0	67.133651	64.602754
MLB-style	1	66.947387	64.751942
MLB-style	2	67.226342	64.345869
GMU-style	0	65.203037	64.976165
GMU-style	1	65.385770	65.326624
GMU-style	2	65.434322	65.444915
Concat-256	0	65.074153	64.723694
Concat-256	1	64.481815	64.319386
Concat-256	2	64.846398	64.671610
Concat-512	0	65.164195	64.932027
Concat-512	1	65.025600	64.774011
Concat-512	2	64.490643	63.894774

Table 24: Own-model VQA projection sensitivities for both models in every comparison. Values are accuracy percentage points, with pointwise 95% cluster-bootstrap intervals.

Comparison	$D_p(f_h)$	$D_p(f_c)$
MLB-style / shared	2.54 [1.98, 3.10]	0.13 [0.01, 0.25]
MLB-style / matched	2.54 [1.98, 3.10]	0.23 [0.12, 0.34]
GMU-style / shared	0.09 [0.00, 0.18]	0.13 [0.01, 0.25]
GMU-style / matched	0.09 [0.00, 0.18]	0.36 [0.21, 0.52]

H REFERENCE SENSITIVITY ACROSS THE NEW TASKS

The reference sweeps use the selected models from Section 4, their fitted seeds and complete scoring sets. For each dataset, two training-reference laws, three draws and three nested sizes give 18 conditions. The initial 256-row reference remains the primary comparison. Ranges and sample SDs across draws quantify empirical reference variation within each law and size. Conditional cluster-bootstrap intervals quantify scoring variation for each realised reference and fixed set of fitted models. These two summaries retain their distinct interpretations. The source-uniform law changes the weighting of the reference population relative to observation-uniform sampling. The schemes also differ in replacement, so their empirical contrast describes both specified sampling designs. For a fixed model comparison, G stays constant across references and $E_p = G - G_p$. The reference-draw SDs of G_p and E_p therefore agree, and their point-estimate ranges have equal widths. Their conditional scoring intervals differ because E_p also retains the paired covariance with the resampled original gap. Each released row includes the reference identifiers, four utilities, paired contrasts and interval endpoints.

Primary-reference summary. At the primary 256-row reference, the new-task study covers four fusion mechanisms in six head/task settings. Four original-gap intervals against matched concatenation are positive, and two span zero. Across 14 partition comparisons, nine projected-gap intervals are positive and five span zero. Excess-sensitivity intervals are positive in one comparison, negative in two and span zero in eleven. The positive case is MLB-style on VQA. The negative cases are GMU-style on VQA and TFN under $V|TA$ on MOSI. These counts describe correlated comparisons under the specified implementations and pointwise inference.

H.1 VQA REFERENCE SENSITIVITY

The sensitivity analysis uses the five configurations and three fitted seeds specified in Section 4. All 11,328 scoring questions and 5,368 source scenes are retained. Eligible references comprise the 18,745 internal training questions from 8,783 scenes. Question-uniform samples contain distinct

Table 25: MLB-style / matched reference sensitivity. Score contrasts use accuracy percentage points. SD is the sample standard deviation over three reference draws. The SD of E_p equals that of G_p because G is fixed.

Law	n	Mean G_p	SD	G_p range	Mean E_p	E_p range
Question	128	0.107	0.083	[0.023, 0.190]	2.195	[2.112, 2.279]
Question	256	0.089	0.096	[-0.005, 0.187]	2.213	[2.115, 2.306]
Question	512	0.178	0.114	[0.052, 0.276]	2.124	[2.026, 2.250]
Scene	128	0.136	0.076	[0.048, 0.188]	2.166	[2.114, 2.254]
Scene	256	0.222	0.059	[0.160, 0.277]	2.080	[2.025, 2.142]
Scene	512	0.171	0.028	[0.148, 0.202]	2.131	[2.100, 2.153]

Table 26: MLB-style / shared reference sensitivity. Score contrasts use accuracy percentage points. SD is the sample standard deviation over three reference draws. The SD of E_p equals that of G_p because G is fixed.

Law	n	Mean G_p	SD	G_p range	Mean E_p	E_p range
Question	128	-0.125	0.099	[-0.216, -0.020]	2.333	[2.228, 2.424]
Question	256	-0.125	0.075	[-0.195, -0.045]	2.333	[2.253, 2.403]
Question	512	-0.012	0.095	[-0.119, 0.063]	2.220	[2.145, 2.327]
Scene	128	-0.068	0.079	[-0.156, -0.003]	2.276	[2.211, 2.363]
Scene	256	0.029	0.050	[-0.023, 0.077]	2.179	[2.131, 2.231]
Scene	512	-0.023	0.004	[-0.026, -0.018]	2.231	[2.226, 2.234]

rows. For draw 8101, the 256-row reference exactly reproduces the original sample. The 128-row reference is nested within it, and 256 additional rows extend it to size 512. The same rule is applied to draws 8102 and 8103. Scene-uniform references use two-stage sampling with replacement: a source scene is sampled uniformly, followed by a uniform question from that scene. Their three sizes are prefixes of a 512-draw sequence. Repeated entries retain their multiplicity. Sampling uses row and scene identities without answer labels. Both models share the actual reference entries. The scene-uniform law changes the reference target because scenes with different question counts receive different weights.

Every condition uses the same 1,000 source-scene bootstrap draws (seed 20260912). A sampled scene contributes all its question rows, including multiplicities, and each bootstrap utility divides by its resampled question count. Utilities are computed separately for each fixed training seed before averaging. Intervals are pointwise 95% linear-percentile intervals for a fixed reference. The released aggregate tables retain every original/projected model utility, all four comparisons and all five signed contrasts. The tables below show means and ranges over the three reference draws.

For MLB-style against matched concatenation, question-uniform projected-gap SDs are 0.083, 0.096 and 0.114 pp at reference sizes 128, 256 and 512. Across these draws, metric variability changes non-monotonically with reference size. All 18 MLB-style projected-gap intervals contain zero for both concatenation baselines, while all corresponding excess-sensitivity intervals are positive. All 18 GMU-style projected-gap intervals are positive for both concatenation baselines. Its excess sensitivity is negative for every interval against matched concatenation, whereas the excess-sensitivity intervals against shared concatenation all contain zero.

Table 27: GMU-style / matched reference sensitivity. Score contrasts use accuracy percentage points. SD is the sample standard deviation over three reference draws. The SD of E_p equals that of G_p because G is fixed.

Law	n	Mean G_p	SD	G_p range	Mean E_p	E_p range
Question	128	0.727	0.028	[0.695, 0.749]	-0.280	[-0.302, -0.248]
Question	256	0.727	0.021	[0.713, 0.751]	-0.279	[-0.303, -0.266]
Question	512	0.735	0.012	[0.722, 0.744]	-0.288	[-0.296, -0.274]
Scene	128	0.739	0.029	[0.710, 0.769]	-0.291	[-0.321, -0.263]
Scene	256	0.735	0.012	[0.722, 0.747]	-0.288	[-0.299, -0.275]
Scene	512	0.729	0.023	[0.703, 0.743]	-0.282	[-0.295, -0.256]

Table 28: GMU-style / shared reference sensitivity. Score contrasts use accuracy percentage points. SD is the sample standard deviation over three reference draws. The SD of E_p equals that of G_p because G is fixed.

Law	n	Mean G_p	SD	G_p range	Mean E_p	E_p range
Question	128	0.486	0.043	[0.445, 0.531]	-0.040	[-0.084, 0.002]
Question	256	0.494	0.009	[0.487, 0.503]	-0.047	[-0.057, -0.041]
Question	512	0.513	0.007	[0.506, 0.518]	-0.067	[-0.072, -0.059]
Scene	128	0.509	0.014	[0.498, 0.525]	-0.062	[-0.079, -0.052]
Scene	256	0.502	0.009	[0.492, 0.511]	-0.056	[-0.065, -0.046]
Scene	512	0.510	0.039	[0.467, 0.544]	-0.064	[-0.097, -0.021]

H.2 Sentiment reference sensitivity

Each dataset retains its five selected configurations, all three training seeds and the three partitions $T|AV$, $A|TV$ and $V|TA$. Reference sizes are 128, 256 and 512, with draws identified by seeds 8101, 8103 and 8104. Segment-uniform sampling first selects 256 training rows without replacement. A uniform 128-row subset supplies the smaller reference, and 256 rows from the complement extend it to 512. The primary 8101/256 reference is preserved exactly. Video-uniform sampling selects a training video uniformly, followed by a uniform segment within that video, with replacement. The 128-row and 256-row references are prefixes of the resulting 512-entry sequence. Repeated segments retain their multiplicity. Both laws use row identifiers and source groups to determine reference membership.

All 18 conditions retain the original test rows and 2,000 shared source-video bootstrap draws with seed 8102. Each resampled video contributes all its segments, and MAE uses the resampled segment count. Utilities are calculated separately for each fixed training seed and then averaged. Pointwise percentile intervals condition on the fitted cohort and each realised reference. The largest cross-pair grid provides exact subgrids for the nested sizes. Recomputed original predictions and the primary projected predictions are checked against their archived counterparts before reuse.

The following tables report every parameter-matched comparison, separating each reference law and size. Means, sample SDs and ranges summarise the three draws, while the direction counts classify their individual conditional intervals. The $+/-/0$ counts indicate intervals above, below and spanning zero. Interval directions use unrounded endpoints. All shared-concatenation comparisons and every condition's four scores and interval endpoints are retained in the released CSV and JSON records.

Table 29: MOSI matched comparisons under segment-uniform references. Three-draw G_p summaries use MAE units. Interval directions follow Section H.2.

Head	Partition	n	Mean $\pm$ SD	Range	$+/-/0$
LMF	$A \mid TV$	128	0.16615 ± 0.00010	$[0.16605, 0.16626]$	3/0/0
LMF	$A \mid TV$	256	0.16609 ± 0.00006	$[0.16603, 0.16615]$	3/0/0
LMF	$A \mid TV$	512	0.16600 ± 0.00001	$[0.16599, 0.16601]$	3/0/0
LMF	$T \mid AV$	128	0.16979 ± 0.00047	$[0.16928, 0.17021]$	3/0/0
LMF	$T \mid AV$	256	0.16954 ± 0.00030	$[0.16923, 0.16984]$	3/0/0
LMF	$T \mid AV$	512	0.16931 ± 0.00011	$[0.16918, 0.16938]$	3/0/0
LMF	$V \mid TA$	128	0.16982 ± 0.00040	$[0.16938, 0.17015]$	3/0/0
LMF	$V \mid TA$	256	0.16963 ± 0.00025	$[0.16937, 0.16987]$	3/0/0
LMF	$V \mid TA$	512	0.16947 ± 0.00009	$[0.16937, 0.16952]$	3/0/0
TFN	$A \mid TV$	128	0.03256 ± 0.00072	$[0.03174, 0.03312]$	1/0/2
TFN	$A \mid TV$	256	0.03221 ± 0.00067	$[0.03146, 0.03277]$	0/0/3
TFN	$A \mid TV$	512	0.03207 ± 0.00029	$[0.03176, 0.03232]$	0/0/3
TFN	$T \mid AV$	128	0.03965 ± 0.00156	$[0.03825, 0.04133]$	3/0/0
TFN	$T \mid AV$	256	0.03950 ± 0.00136	$[0.03837, 0.04101]$	3/0/0
TFN	$T \mid AV$	512	0.03920 ± 0.00053	$[0.03871, 0.03976]$	3/0/0
TFN	$V \mid TA$	128	0.04241 ± 0.00087	$[0.04187, 0.04341]$	3/0/0
TFN	$V \mid TA$	256	0.04263 ± 0.00079	$[0.04198, 0.04351]$	3/0/0
TFN	$V \mid TA$	512	0.04257 ± 0.00028	$[0.04231, 0.04288]$	3/0/0

Table 30: MOSI matched comparisons under segment-uniform references. Three-draw E_p summaries use MAE units. Interval directions follow Section H.2.

Head	Partition	n	Mean $\pm$ SD	Range	$+/-/0$
LMF	$A \mid TV$	128	-0.00043 ± 0.00010	$[-0.00054, -0.00033]$	0/0/3
LMF	$A \mid TV$	256	-0.00036 ± 0.00006	$[-0.00042, -0.00031]$	0/0/3
LMF	$A \mid TV$	512	-0.00027 ± 0.00001	$[-0.00028, -0.00026]$	0/0/3
LMF	$T \mid AV$	128	-0.00407 ± 0.00047	$[-0.00448, -0.00356]$	0/0/3
LMF	$T \mid AV$	256	-0.00381 ± 0.00030	$[-0.00412, -0.00351]$	0/0/3
LMF	$T \mid AV$	512	-0.00359 ± 0.00011	$[-0.00366, -0.00346]$	0/0/3
LMF	$V \mid TA$	128	-0.00409 ± 0.00040	$[-0.00443, -0.00365]$	0/0/3
LMF	$V \mid TA$	256	-0.00390 ± 0.00025	$[-0.00414, -0.00364]$	0/0/3
LMF	$V \mid TA$	512	-0.00375 ± 0.00009	$[-0.00380, -0.00365]$	0/0/3
TFN	$A \mid TV$	128	0.00190 ± 0.00072	$[0.00134, 0.00272]$	0/0/3
TFN	$A \mid TV$	256	0.00225 ± 0.00067	$[0.00170, 0.00300]$	0/0/3
TFN	$A \mid TV$	512	0.00239 ± 0.00029	$[0.00214, 0.00270]$	0/0/3
TFN	$T \mid AV$	128	-0.00519 ± 0.00156	$[-0.00686, -0.00379]$	0/0/3
TFN	$T \mid AV$	256	-0.00503 ± 0.00136	$[-0.00654, -0.00391]$	0/0/3
TFN	$T \mid AV$	512	-0.00474 ± 0.00053	$[-0.00530, -0.00424]$	0/0/3
TFN	$V \mid TA$	128	-0.00794 ± 0.00087	$[-0.00894, -0.00740]$	0/1/2
TFN	$V \mid TA$	256	-0.00816 ± 0.00079	$[-0.00904, -0.00751]$	0/1/2
TFN	$V \mid TA$	512	-0.00811 ± 0.00028	$[-0.00841, -0.00785]$	0/1/2

Table 31: MOSI matched comparisons under video-uniform references. Three-draw G_p summaries use MAE units. Interval directions follow Section H.2.

Head	Partition	n	Mean $\pm$ SD	Range	$+/-/0$
LMF	$A \mid TV$	128	0.16592 ± 0.00002	$[0.16591, 0.16595]$	3/0/0
LMF	$A \mid TV$	256	0.16599 ± 0.00014	$[0.16590, 0.16616]$	3/0/0
LMF	$A \mid TV$	512	0.16601 ± 0.00012	$[0.16587, 0.16610]$	3/0/0
LMF	$T \mid AV$	128	0.16875 ± 0.00020	$[0.16862, 0.16897]$	3/0/0
LMF	$T \mid AV$	256	0.16899 ± 0.00058	$[0.16862, 0.16966]$	3/0/0
LMF	$T \mid AV$	512	0.16905 ± 0.00038	$[0.16861, 0.16928]$	3/0/0
LMF	$V \mid TA$	128	0.16912 ± 0.00016	$[0.16901, 0.16931]$	3/0/0
LMF	$V \mid TA$	256	0.16927 ± 0.00037	$[0.16904, 0.16969]$	3/0/0
LMF	$V \mid TA$	512	0.16928 ± 0.00023	$[0.16902, 0.16944]$	3/0/0
TFN	$A \mid TV$	128	0.03237 ± 0.00055	$[0.03203, 0.03301]$	0/0/3
TFN	$A \mid TV$	256	0.03238 ± 0.00031	$[0.03210, 0.03272]$	0/0/3
TFN	$A \mid TV$	512	0.03220 ± 0.00016	$[0.03207, 0.03238]$	0/0/3
TFN	$T \mid AV$	128	0.03857 ± 0.00069	$[0.03778, 0.03906]$	3/0/0
TFN	$T \mid AV$	256	0.03934 ± 0.00061	$[0.03881, 0.04001]$	3/0/0
TFN	$T \mid AV$	512	0.03890 ± 0.00048	$[0.03851, 0.03943]$	3/0/0
TFN	$V \mid TA$	128	0.04208 ± 0.00009	$[0.04201, 0.04218]$	3/0/0
TFN	$V \mid TA$	256	0.04253 ± 0.00011	$[0.04244, 0.04265]$	3/0/0
TFN	$V \mid TA$	512	0.04231 ± 0.00019	$[0.04212, 0.04249]$	3/0/0

Table 32: MOSI matched comparisons under video-uniform references. Three-draw E_p summaries use MAE units. Interval directions follow Section H.2.

Head	Partition	n	Mean $\pm$ SD	Range	$+/-/0$
LMF	$A \mid TV$	128	-0.00020 ± 0.00002	$[-0.00022, -0.00019]$	0/0/3
LMF	$A \mid TV$	256	-0.00026 ± 0.00014	$[-0.00043, -0.00017]$	0/0/3
LMF	$A \mid TV$	512	-0.00029 ± 0.00012	$[-0.00038, -0.00015]$	0/0/3
LMF	$T \mid AV$	128	-0.00302 ± 0.00020	$[-0.00324, -0.00289]$	0/0/3
LMF	$T \mid AV$	256	-0.00326 ± 0.00058	$[-0.00393, -0.00289]$	0/0/3
LMF	$T \mid AV$	512	-0.00332 ± 0.00038	$[-0.00356, -0.00288]$	0/0/3
LMF	$V \mid TA$	128	-0.00339 ± 0.00016	$[-0.00358, -0.00329]$	0/0/3
LMF	$V \mid TA$	256	-0.00354 ± 0.00037	$[-0.00397, -0.00331]$	0/0/3
LMF	$V \mid TA$	512	-0.00355 ± 0.00023	$[-0.00371, -0.00329]$	0/0/3
TFN	$A \mid TV$	128	0.00209 ± 0.00055	$[0.00146, 0.00244]$	0/0/3
TFN	$A \mid TV$	256	0.00209 ± 0.00031	$[0.00175, 0.00237]$	0/0/3
TFN	$A \mid TV$	512	0.00226 ± 0.00016	$[0.00209, 0.00239]$	0/0/3
TFN	$T \mid AV$	128	-0.00410 ± 0.00069	$[-0.00460, -0.00331]$	0/0/3
TFN	$T \mid AV$	256	-0.00488 ± 0.00061	$[-0.00555, -0.00435]$	0/0/3
TFN	$T \mid AV$	512	-0.00444 ± 0.00048	$[-0.00497, -0.00404]$	0/0/3
TFN	$V \mid TA$	128	-0.00761 ± 0.00009	$[-0.00771, -0.00754]$	0/0/3
TFN	$V \mid TA$	256	-0.00807 ± 0.00011	$[-0.00818, -0.00797]$	0/0/3
TFN	$V \mid TA$	512	-0.00785 ± 0.00019	$[-0.00802, -0.00765]$	0/0/3

Table 33: MOSEI matched comparisons under segment-uniform references. Three-draw G_p summaries use MAE units. Interval directions follow Section H.2.

Head	Partition	n	Mean ± SD	Range	$+/-/0$
LMF	$A \mid TV$	128	0.10081 ± 0.00000	$[0.10081, 0.10081]$	3/0/0
LMF	$A \mid TV$	256	0.10086 ± 0.00002	$[0.10083, 0.10087]$	3/0/0
LMF	$A \mid TV$	512	0.10084 ± 0.00004	$[0.10081, 0.10089]$	3/0/0
LMF	$T \mid AV$	128	0.10061 ± 0.00044	$[0.10025, 0.10109]$	3/0/0
LMF	$T \mid AV$	256	0.10081 ± 0.00025	$[0.10063, 0.10110]$	3/0/0
LMF	$T \mid AV$	512	0.10079 ± 0.00034	$[0.10058, 0.10118]$	3/0/0
LMF	$V \mid TA$	128	0.10049 ± 0.00043	$[0.10013, 0.10096]$	3/0/0
LMF	$V \mid TA$	256	0.10064 ± 0.00023	$[0.10048, 0.10091]$	3/0/0
LMF	$V \mid TA$	512	0.10063 ± 0.00030	$[0.10045, 0.10098]$	3/0/0
TFN	$A \mid TV$	128	-0.00074 ± 0.00011	$[-0.00083, -0.00061]$	0/0/3
TFN	$A \mid TV$	256	-0.00051 ± 0.00005	$[-0.00054, -0.00045]$	0/0/3
TFN	$A \mid TV$	512	-0.00048 ± 0.00006	$[-0.00054, -0.00043]$	0/0/3
TFN	$T \mid AV$	128	-0.00081 ± 0.00005	$[-0.00087, -0.00077]$	0/0/3
TFN	$T \mid AV$	256	-0.00049 ± 0.00013	$[-0.00063, -0.00037]$	0/0/3
TFN	$T \mid AV$	512	-0.00055 ± 0.00024	$[-0.00081, -0.00036]$	0/0/3
TFN	$V \mid TA$	128	-0.00102 ± 0.00010	$[-0.00109, -0.00090]$	0/0/3
TFN	$V \mid TA$	256	-0.00095 ± 0.00016	$[-0.00113, -0.00086]$	0/0/3
TFN	$V \mid TA$	512	-0.00097 ± 0.00019	$[-0.00118, -0.00082]$	0/0/3

Table 34: MOSEI matched comparisons under segment-uniform references. Three-draw E_p summaries use MAE units. Interval directions follow Section H.2.

Head	Partition	n	Mean ± SD	Range	$+/-/0$
LMF	$A \mid TV$	128	-0.00027 ± 0.00000	$[-0.00027, -0.00027]$	0/0/3
LMF	$A \mid TV$	256	-0.00031 ± 0.00002	$[-0.00033, -0.00028]$	0/0/3
LMF	$A \mid TV$	512	-0.00029 ± 0.00004	$[-0.00034, -0.00026]$	0/0/3
LMF	$T \mid AV$	128	-0.00006 ± 0.00044	$[-0.00055, 0.00030]$	0/0/3
LMF	$T \mid AV$	256	-0.00027 ± 0.00025	$[-0.00055, -0.00009]$	0/0/3
LMF	$T \mid AV$	512	-0.00024 ± 0.00034	$[-0.00063, -0.00004]$	0/0/3
LMF	$V \mid TA$	128	0.00006 ± 0.00043	$[-0.00042, 0.00042]$	0/0/3
LMF	$V \mid TA$	256	-0.00010 ± 0.00023	$[-0.00036, 0.00006]$	0/0/3
LMF	$V \mid TA$	512	-0.00008 ± 0.00030	$[-0.00043, 0.00009]$	0/0/3
TFN	$A \mid TV$	128	-0.00006 ± 0.00011	$[-0.00019, 0.00003]$	0/0/3
TFN	$A \mid TV$	256	-0.00029 ± 0.00005	$[-0.00035, -0.00026]$	0/0/3
TFN	$A \mid TV$	512	-0.00032 ± 0.00006	$[-0.00036, -0.00026]$	0/0/3
TFN	$T \mid AV$	128	0.00001 ± 0.00005	$[-0.00003, 0.00007]$	0/0/3
TFN	$T \mid AV$	256	-0.00031 ± 0.00013	$[-0.00043, -0.00017]$	0/0/3
TFN	$T \mid AV$	512	-0.00025 ± 0.00024	$[-0.00044, 0.00001]$	0/0/3
TFN	$V \mid TA$	128	0.00022 ± 0.00010	$[0.00010, 0.00029]$	0/0/3
TFN	$V \mid TA$	256	0.00015 ± 0.00016	$[0.00006, 0.00033]$	0/0/3
TFN	$V \mid TA$	512	0.00017 ± 0.00019	$[0.00002, 0.00038]$	0/0/3

Table 35: MOSEI matched comparisons under video-uniform references. Three-draw G_p summaries use MAE units. Interval directions follow Section H.2.

Head	Partition	n	Mean ± SD	Range	$+/-/0$
LMF	$A \mid TV$	128	0.10095 ± 0.00009	$[0.10085, 0.10101]$	3/0/0
LMF	$A \mid TV$	256	0.10094 ± 0.00006	$[0.10087, 0.10097]$	3/0/0
LMF	$A \mid TV$	512	0.10094 ± 0.00001	$[0.10093, 0.10095]$	3/0/0
LMF	$T \mid AV$	128	0.10134 ± 0.00056	$[0.10074, 0.10185]$	3/0/0
LMF	$T \mid AV$	256	0.10145 ± 0.00032	$[0.10110, 0.10174]$	3/0/0
LMF	$T \mid AV$	512	0.10156 ± 0.00008	$[0.10147, 0.10163]$	3/0/0
LMF	$V \mid TA$	128	0.10111 ± 0.00049	$[0.10059, 0.10157]$	3/0/0
LMF	$V \mid TA$	256	0.10122 ± 0.00027	$[0.10094, 0.10147]$	3/0/0
LMF	$V \mid TA$	512	0.10132 ± 0.00008	$[0.10124, 0.10139]$	3/0/0
TFN	$A \mid TV$	128	-0.00061 ± 0.00019	$[-0.00079, -0.00041]$	0/0/3
TFN	$A \mid TV$	256	-0.00082 ± 0.00018	$[-0.00103, -0.00071]$	0/0/3
TFN	$A \mid TV$	512	-0.00074 ± 0.00013	$[-0.00088, -0.00061]$	0/0/3
TFN	$T \mid AV$	128	-0.00001 ± 0.00056	$[-0.00064, 0.00043]$	0/0/3
TFN	$T \mid AV$	256	-0.00027 ± 0.00027	$[-0.00055, -0.00001]$	0/0/3
TFN	$T \mid AV$	512	-0.00013 ± 0.00006	$[-0.00018, -0.00006]$	0/0/3
TFN	$V \mid TA$	128	-0.00048 ± 0.00049	$[-0.00104, -0.00012]$	0/0/3
TFN	$V \mid TA$	256	-0.00050 ± 0.00026	$[-0.00080, -0.00035]$	0/0/3
TFN	$V \mid TA$	512	-0.00044 ± 0.00009	$[-0.00054, -0.00038]$	0/0/3

Table 36: MOSEI matched comparisons under video-uniform references. Three-draw E_p summaries use MAE units. Interval directions follow Section H.2.

Head	Partition	n	Mean ± SD	Range	$+/-/0$
LMF	$A \mid TV$	128	-0.00040 ± 0.00009	$[-0.00046, -0.00030]$	0/0/3
LMF	$A \mid TV$	256	-0.00039 ± 0.00006	$[-0.00043, -0.00033]$	0/0/3
LMF	$A \mid TV$	512	-0.00040 ± 0.00001	$[-0.00041, -0.00038]$	0/0/3
LMF	$T \mid AV$	128	-0.00080 ± 0.00056	$[-0.00130, -0.00020]$	0/0/3
LMF	$T \mid AV$	256	-0.00091 ± 0.00032	$[-0.00119, -0.00056]$	0/0/3
LMF	$T \mid AV$	512	-0.00101 ± 0.00008	$[-0.00108, -0.00092]$	0/0/3
LMF	$V \mid TA$	128	-0.00057 ± 0.00049	$[-0.00102, -0.00004]$	0/0/3
LMF	$V \mid TA$	256	-0.00068 ± 0.00027	$[-0.00093, -0.00039]$	0/0/3
LMF	$V \mid TA$	512	-0.00078 ± 0.00008	$[-0.00085, -0.00069]$	0/0/3
TFN	$A \mid TV$	128	-0.00019 ± 0.00019	$[-0.00039, -0.00001]$	0/0/3
TFN	$A \mid TV$	256	0.00002 ± 0.00018	$[-0.00009, 0.00023]$	0/0/3
TFN	$A \mid TV$	512	-0.00006 ± 0.00013	$[-0.00019, 0.00008]$	0/0/3
TFN	$T \mid AV$	128	-0.00079 ± 0.00056	$[-0.00123, -0.00016]$	0/0/3
TFN	$T \mid AV$	256	-0.00053 ± 0.00027	$[-0.00079, -0.00025]$	0/0/3
TFN	$T \mid AV$	512	-0.00067 ± 0.00006	$[-0.00074, -0.00062]$	0/0/3
TFN	$V \mid TA$	128	-0.00032 ± 0.00049	$[-0.00068, 0.00024]$	0/0/3
TFN	$V \mid TA$	256	-0.00029 ± 0.00026	$[-0.00045, -0.00000]$	0/0/3
TFN	$V \mid TA$	512	-0.00036 ± 0.00009	$[-0.00042, -0.00026]$	0/0/3

Table 37: MOSI comparisons with shared concatenation across all 18 reference conditions. Point-estimate ranges use MAE units. Direction counts classify each conditional interval using unrounded endpoints.

Head	Partition	G_p range	$+/-/0$	E_p range	$+/-/0$
LMF	$T \mid AV$	$[-0.17396, -0.17186]$	0/18/0	$[-0.00498, -0.00288]$	0/0/18
LMF	$A \mid TV$	$[-0.17688, -0.17616]$	0/18/0	$[-0.00068, 0.00004]$	0/0/18
LMF	$V \mid TA$	$[-0.17339, -0.17208]$	0/18/0	$[-0.00476, -0.00345]$	0/0/18
TFN	$T \mid AV$	$[0.00892, 0.01294]$	0/0/18	$[-0.00759, -0.00356]$	0/0/18
TFN	$A \mid TV$	$[0.00227, 0.00400]$	0/0/18	$[0.00136, 0.00309]$	0/0/18
TFN	$V \mid TA$	$[0.01366, 0.01534]$	0/0/18	$[-0.00998, -0.00830]$	0/14/4

Table 38: MOSEI comparisons with shared concatenation across all 18 reference conditions. Point-estimate ranges use MAE units. Direction counts classify each conditional interval using unrounded endpoints.

Head	Partition	G_p range	$+/-/0$	E_p range	$+/-/0$
LMF	$T \mid AV$	$[0.01401, 0.01735]$	18/0/0	$[-0.00626, -0.00292]$	0/16/2
LMF	$A \mid TV$	$[0.01376, 0.01613]$	18/0/0	$[-0.00504, -0.00267]$	0/18/0
LMF	$V \mid TA$	$[0.01291, 0.01480]$	18/0/0	$[-0.00372, -0.00182]$	0/6/12
TFN	$T \mid AV$	$[-0.00265, 0.00016]$	0/0/18	$[-0.00243, 0.00038]$	0/0/18
TFN	$A \mid TV$	$[-0.00070, 0.00145]$	0/0/18	$[-0.00372, -0.00157]$	0/9/9
TFN	$V \mid TA$	$[-0.00334, -0.00150]$	0/0/18	$[-0.00077, 0.00107]$	0/0/18

I Computational cost and reuse

For m scoring rows and n separate reference rows, exact projection requires $2mn + n^2$ cross-pair evaluations per predictor and partition. Ordinary scoring requires m original-input evaluations. When the scoring and reference sets coincide, one $m \times m$ grid supplies the row means, column means and grand mean. Independent unimodal transformations can be evaluated once and reused across cross-pairs. Joint nonlinear operations remain in the cross-pair computation. Thus the cross-pair count and the ratio of measured wall times describe different costs.

Table 39: Exact cross-pair counts per predictor and partition. The separate-reference columns use $n = 256$. Ordinary scoring evaluates m original rows. The full-test MM-IMDb grid uses $60,824,401$ pairs.

Dataset	m	Separate-reference pairs	Pairs / m
MM-IMDb	7,799	4,058,624	520.4
MOSI	686	416,768	607.5
MOSEI	4,659	2,450,944	526.1
VQA	11,328	5,865,472	517.8

Movie timing records. The 256-row movie reference analysis retains 36 measured projection records across two banks, two models, three seeds and three partitions. They total 538.70 seconds. For w2v, median projection times are 1.64 seconds for matched concatenation and 30.14 seconds for Mamba-style. For ST, the corresponding medians are 1.03 and 27.06 seconds. Each median summarises nine model-seed/partition runs. These archived times cover cached projection and result transfer, with accelerator synchronisation before timing ends. Feature acquisition, training and initial cache construction lie outside that interval. The numerical pair-count comparison above supplies the corresponding ordinary-evaluation reference.

VQA timing procedure. The timing study reuses the stored image and question representations on an RTX 5090 workstation. Ordinary scoring includes the fusion model's input transformations.

Table 40: Observed VQA fusion-level diagnostic cost for seed 0 on a shared RTX 5090 workstation. Times are milliseconds. Standard scoring evaluates all 11,328 questions. Each incremental diagnostic time includes reference/scoring tensor assembly, separated input-transform caching and exact projection. Scoring and projection use a warm-up and medians of three synchronized repetitions. Preparation is measured once. The final column is the incremental $n = 256$ cost divided by standard scoring time.

Model	Scoring	$n = 128$	$n = 256$	$n = 512$	Ratio
Concat-128	6.09	235.6	240.8	260.8	39.5
MLB-style	3.25	229.0	203.5	209.2	62.5
GMU-style	5.75	332.6	308.1	346.2	53.6
Concat-256	3.11	245.6	245.7	247.1	79.1
Concat-512	3.59	274.8	239.9	275.0	66.8

The incremental diagnostic includes scoring/reference tensor assembly, caching of independent input transformations and exact projection. Scoring and projection each use a warm-up and three synchronised measurements, summarised by their median. Preparation is measured once for each condition. The workstation is shared, so small timing differences across sizes include execution variability and fixed overhead. Raw image feature extraction and question-vector fitting are outside these fusion-level timings. The released records retain stage-level times and original/projected output checks. Loading, standardising and transferring the feature arrays takes 0.362 seconds. Generating common bootstrap draws takes 0.062 seconds. The CPU statistical summary across all 18 conditions takes 0.196 seconds.

Sentiment timing procedure. Ordinary scoring includes the trained modality subnetworks and fusion readout on stored benchmark features. Each original-scoring time is a single warmed measurement for the specified fitted seed. Independent projection timings use cached unimodal transformations, with a warm-up and three synchronised repetitions. They fix seed 0, the text partition and segment-uniform draw 8101, and evaluate each reference size independently. The projection/scoring ratio below compares these specified stages. Scoring-set and reference-union encoding times are listed separately. The latter union supports all 18 reference conditions and is encoded once per fit. CPU summary times include auxiliary metrics, paired video bootstrap and aggregate-file output. Input loading, feature acquisition and training are separate from the reported scoring and projection stages. All timings use the shared RTX 5090 workstation.

Table 41: Sentiment fusion-level times in seconds. Original scoring uses one warmed seed-0 measurement. Projection entries are medians of three repetitions at each size. The final column divides cached projection time at $n = 256$ by original-scoring time.

Model	Original	$n = 128$	$n = 256$	$n = 512$	Ratio
MOSI					
LMF	0.087	0.186	0.433	0.976	5.0
LMF matched concat.	0.061	0.186	0.399	0.935	6.5
TFN	0.066	0.401	0.869	1.985	13.3
TFN matched concat.	0.068	0.259	0.537	1.284	7.9
Shared concat.	0.072	0.263	0.546	1.264	7.6
MOSEI					
LMF	0.552	0.998	2.018	4.205	3.7
LMF matched concat.	0.331	0.518	1.051	2.147	3.2
TFN	0.345	2.789	5.664	11.663	16.4
TFN matched concat.	0.436	0.840	1.700	3.478	3.9
Shared concat.	0.321	0.832	1.673	3.446	5.2

The complete 18-condition MOSI CPU statistical summary takes 0.399 seconds. The complete 18-condition MOSEI CPU statistical summary takes 1.444 seconds.

Table 42: One-time sentiment encoding costs in seconds for seed 0. The reference union contains the distinct training rows required by all 18 conditions.

Dataset	Model	Scoring-set encoding	Reference-union encoding
MOSI	LMF	0.050	0.123
MOSI	LMF matched concat.	0.054	0.094
MOSI	TFN	0.059	0.092
MOSI	TFN matched concat.	0.060	0.089
MOSI	Shared concat.	0.073	0.122
MOSEI	LMF	0.365	0.310
MOSEI	LMF matched concat.	0.326	0.194
MOSEI	TFN	0.324	0.192
MOSEI	TFN matched concat.	0.418	0.259
MOSEI	Shared concat.	0.319	0.180

J HISTORICALLY INFORMED MOVIE COMPARISON

Fixed-checkpoint study. We specified the targeted protocol after reviewing the historical test results and selected Mamba-style because it led the movie-head comparison. The experiment uses both movie banks, seeds 0, 1 and 2, and the four models defined above. Eighteen baseline checkpoints were reused after training-bank, validation-bank and weight hashes matched. Six Mamba-style models were trained once using the fixed recipe.

Each fitted model supplies all three partitions, with identical original logits across partitions. Projection evaluates all $7,799^2$ test cross-pairs using float32 forward passes and float64 accumulation. Cached input transformations were checked against the original forward function. The study contains 54 full-grid evaluations and 18 analytically invariant additive records, each checked on a small grid. Independent AUC recomputation agreed within 3.33×10^{-16}. Section N provides the complete numerical and execution checks.

Paired uncertainty and reference sensitivity. For each bank, 1,000 bootstrap draws resample the 7,799 movie rows. Each draw is shared across models, partitions and the three fitted seeds. The projection remains fixed, and the draw reweights the saved predictions when computing macro-AUC. Only bootstrap draws containing both outcome classes in every genre enter the analysis. Both banks retained all 1,000 draws. Pointwise 95% percentile intervals summarise the mean of the three seed-specific contrasts. They condition on fitted models, seeds and the full-test reference. Reference sensitivity is evaluated separately using subsets of 256 and 1,024 movies, with three prespecified draws at each size. These subsets share the scored test pool and can overlap.

J.1 FIXED-CHECKPOINT COMPARISON WITH A SELECTED NONLINEAR BASELINE

Table 43 reports the fixed-checkpoint scores and parameter counts underlying the MM-IMDb comparisons in Section 5.2. Table 44 gives the corresponding partition-specific G, G_p and E_p estimates.

On ST, selected concatenation improved original AUC over default concatenation by 0.178 pp, with a 95% interval of [0.128, 0.237]. Its three projected-score improvements also had intervals above zero. Figure 5 compares the gaps against default and selected concatenation.

Reference sensitivity. For Mamba-style versus selected concatenation, six reference subsets changed the mean projected gaps by at most 0.017 pp on w2v and 0.007 pp on ST. All 36 subset-partition means retained their full-reference signs. The w2v excess projection sensitivity under $a \mid bc$ changed sign in two subsets against default concatenation and one against selected concatenation. These descriptive checks quantify reference sensitivity within the same movie pool. Section N reports the full subset ranges and own-model sensitivities.

Table 43: Original fixed-checkpoint scores and model sizes. Macro-AUC is mean ± sample SD across three seeds. Selected concatenation uses six validation configurations per bank, each trained with three seeds. The other models use one configuration. Each listed model is held fixed across all projection partitions. Default concatenation reuses separately trained additional-baseline checkpoints (Appendix M). Its training implementation differs from historical shared-recipe concatenation in Table 57, including layer-initialisation order.

Bank	Model	Parameters	Macro-AUC
w2v	Default concat	356,375	0.8539 ± 0.0006
w2v	Selected concat	712,727	0.8567 ± 0.0006
w2v	Strictly additive	368,709	0.8573 ± 0.0008
w2v	Mamba-style	2,308,119	0.8710 ± 0.0008
ST	Default concat	476,183	0.9314 ± 0.0004
ST	Selected concat	952,343	0.9332 ± 0.0002
ST	Strictly additive	488,517	0.9323 ± 0.0004
ST	Mamba-style	2,427,927	0.9344 ± 0.0007

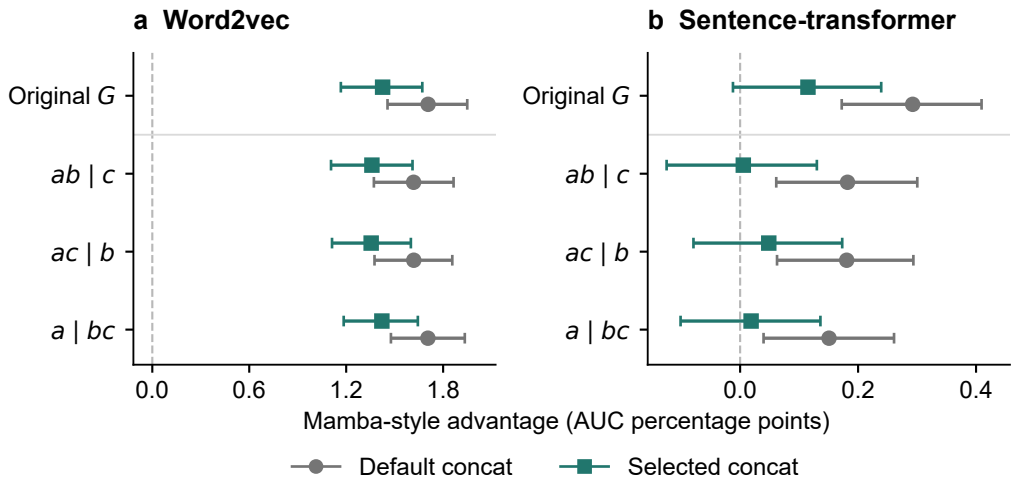

Figure 5: Mamba-style relative to default and selected concatenation. The first row shows G, and the remaining rows show G_p for the indicated partitions. Points show means across three seed-specific contrasts. Bars show pointwise 95% row-reweighting intervals from 1,000 paired resamples of the 7,799 test movies. Fitted models, seeds and the full-test reference remain fixed. Axis units are AUC percentage points, and panel ranges differ.

Table 44: Fixed-checkpoint Mamba-style versus validation-selected concatenation. Entries are three-seed mean differences with pointwise 95% paired row-reweighting intervals (AUC percentage points). The same fitted models are used for every partition. Intervals condition on these fits and the full-test projection reference.

Text	Partition	G	G_p	E_p
w2v	$ab \mid c$	1.43 $[1.17, 1.67]$	1.36 $[1.11, 1.61]$	0.07 $[-0.01, 0.14]$
w2v	$ac \mid b$	1.43 $[1.17, 1.67]$	1.35 $[1.11, 1.60]$	0.07 $[-0.02, 0.16]$
w2v	$a \mid bc$	1.43 $[1.17, 1.67]$	1.42 $[1.18, 1.64]$	0.00 $[-0.09, 0.09]$
ST	$ab \mid c$	0.12 $[-0.01, 0.24]$	0.01 $[-0.12, 0.13]$	0.11 $[0.07, 0.15]$
ST	$ac \mid b$	0.12 $[-0.01, 0.24]$	0.05 $[-0.08, 0.17]$	0.07 $[0.02, 0.11]$
ST	$a \mid bc$	0.12 $[-0.01, 0.24]$	0.02 $[-0.10, 0.14]$	0.10 $[0.05, 0.15]$

K Movie Data and Implementation Details

The 44 metadata features include year, runtime, aspect ratio, US certificate level, and counts of production companies, countries and languages. Selected country, language, colour and sound-format indicators complete the block. These features describe production and release information, while the target comprises the 23 most frequent genres. All 25,959 poster files decode, and both text banks preserve the same movie ordering.

Metadata fitting and text-source audit. Movie identifiers are unique within each split, and all three splits are disjoint. Both banks have identical identifier arrays and ordering, matching the local dataset split file exactly. The dataset's `dev` split supplies validation observations for all movie experiments.

The historical metadata script selected frequent categories using the complete archive. Refitting on the 15,552 training movies retained identical country, language, colour and sound categories. Only two country columns changed order, requiring alignment by feature name. After alignment, the complete $25,959 \times 44$ pre-standardisation matrix and all six standardised metadata blocks exactly matched the archived arrays. This equivalence supports reuse of the original banks and checkpoints. The supplied train-only procedure includes the column mapping required when applying existing weights. The metadata "other" indicator equals one when no retained category is present.

Both text pipelines use identical plot or plot-outline strings after stripping. The sentence-transformer script's additional title fallback triggers zero times across all splits. The local `all-mpnet-base-v2` configuration uses a 384-token maximum, mean-token pooling and a final normalisation module. This internal normalisation remains active with the builder's `normalize_embeddings=False` setting. The raw stored sentence embeddings have unit norm within 1.2×10^{-7}. The inspected local cache revision is `e8c3b32edf5434bc2275fc9bab85f82640a19130`. The model card is available at https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

L Movie Feature Construction and Head Implementations

Word-vector features. The w2v pipeline reads the dataset-provided `metadata.npy` lookup and word-to-index dictionary. The inspected file contains 41,611 float32 vectors of dimension 300 and a 69,980-entry dictionary. The builder concatenates plot strings and falls back to the plot outline. After lowercasing, it retains matches to `[a-z']+` and individual punctuation tokens from `.,!?;:`. It averages vectors for all covered token occurrences, including repeated occurrences. A token must occur in the dictionary and have a lookup row to contribute. The entire text is processed, with equal weighting and arithmetic-mean pooling. A plot with zero covered tokens produces an all-zero

vector. Stored metadata reports 97.66% token-occurrence coverage under this w2v tokeniser. The build records identify the local lookup artifact used for feature construction.

Poster features. The builder converts each decoded JPEG to RGB and resizes it directly to 224×224, changing aspect ratio when necessary. This fixed transform uses the full image for deterministic feature extraction. After conversion to a tensor, channels are normalised by means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$. The encoder is torchvision DenseNet-121 with `DenseNet121_Weights.IMAGENET1K_V1`. Its feature module runs in evaluation mode, followed by ReLU and adaptive spatial-average pooling to a 1,024-dimensional vector. All 25,959 posters decoded successfully and are marked available in both banks.

In the inspected environment, the current implementation uses bilinear interpolation with antialiasing. The poster-cache reuse check verifies the ordered movie identifiers. Archived feature-bank hashes identify the exact inputs used in the reported fixed-bank comparison.

Text embeddings and scaling. The alternate builder uses `SentenceTransformer` to encode plots with `all-mpnet-base-v2`. It uses batch size 64, NumPy output and `normalize_embeddings=False`, then casts to float32. It copies the poster, metadata, label, availability and identifier arrays directly from the source bank and replaces only the stored text array. It allows an additional title fallback, which the text-source audit found unused in this collection. The text-source audit in Section K details the internal normalisation, token cap and model revision. The two text pipelines differ in pretraining, tokenisation, truncation, pooling and dimensionality. For each bank, feature coordinates are centred and scaled using float64 training-set population means and standard deviations. The resulting standardised arrays are then cast to float32 for model execution. A training standard deviation below 10^{-8} is replaced by 1 before scaling.

Specific frozen-bank implementations. Table 45 describes the nine implementations evaluated on the frozen feature banks. The modality encoders are identities, and each final readout produces 23 logits. Unless stated otherwise, projected token width is 256 and attention uses four heads with attention-weight dropout 0.1. The final hidden ReLU layer has width 256 in these implementations. Readout dropout is 0.3 before the first linear layer and 0.2 before the output layer. Parameter totals are constructed on CPU for the actual input dimensions and include registered parameters even when disconnected from the task loss.

CGFM and CGFM-CA use additional dropout 0.1 before their final output. The Perceiver FFNs also use dropout 0.1 within each block. CA+projection includes 197,376 auxiliary parameters outside the evaluated logit path. Its construction consumes random numbers before later layers are initialised, producing a different initialisation sequence from PPA under the same seed. It is evaluated with the auxiliary loss inactive throughout training. TRACE and CGFM-CA use two keys per query, retaining query-dependent attention.

Table 45: Concrete frozen-bank heads. Nominal parameter totals are in millions for w2v/ST. Exact integer counts and source hashes are provided in the release. All implementations use the shared training recipe, with the capacities shown below.

Head	Implemented structure	Parameters w2v/ST
Concat	Raw block concatenation followed by a hidden 256-wide ReLU MLP.	0.356/0.476
PPA	Six directed attention modules with one key–value token each and query-independent evaluation weights. Per-block residual sums and LayerNorm precede a joint MLP.	2.134/2.254
Bilinear	Three rank-8 pairwise products with rank averaging, LayerNorm and ReLU. Their outputs join a nonlinear 256-wide ReLU shortcut in the final MLP.	6.236/8.273
Perceiver-style	Eight learned latent queries attend to three modality tokens. Two cross-attention/self-attention/FFN blocks use FFN width 1024, followed by latent-query mean pooling.	2.534/2.654
Mamba-style	Three position-encoded tokens feed two selective-state-space blocks per direction, with state 16, expansion 2 and convolution width 4. Aligned directions are summed, followed by LayerNorm, flattening and the MLP.	2.308/2.428
TRACE	Three modules each query the other two modality tokens as a length-two key–value sequence. Residual connections and LayerNorm precede the joint MLP.	1.345/1.465
CGFM	Metadata-conditioned 128-wide FiLM generators modulate text/poster projections. Their outputs join the elementwise product and unprojected metadata in a hidden 512/256 MLP.	1.037/1.156
CGFM-CA	Metadata FiLM precedes two attention modules. Each queries the other modulated block and projected metadata as two keys before a hidden 512/256 MLP.	1.444/1.564
CA+projection	PPA logit graph with two additional 256–256–128 projection MLPs. These auxiliary modules remain outside the evaluated task-loss and logit paths.	2.332/2.451

M Baseline Implementation and Auxiliary Metrics

Optimisation. The additive and selected concatenation experiments use the two standardised movie banks described in Section 4. Each configuration uses seeds 0, 1, and 2 with mean binary cross-entropy and constant learning rate throughout 30 epochs. The checkpoint with the greatest validation macro-AUC is retained for each run. The three-seed mean validation score then selects one concatenation configuration per bank. Exact ties retain the earliest checkpoint or the first configuration identifier.

Execution. The additional baseline runner directly initialises its classifier with 23 outputs. It uses float32, disables TF32, and enables deterministic algorithms. The shared-recipe runner initialises a scalar classifier before expanding its output layer. The two initialisation procedures consequently consume random numbers in different orders. Run records preserve configurations, selected epochs, validation scores, hashes, and aggregate test metrics. The study completed 42 training runs across two banks and three seeds. Each bank used six concatenation configurations and one additive configuration. Validation configuration selection preceded all 12 test evaluations of selected concatenation and additive models.

Micro-F1. A global threshold is selected by maximising validation micro-F1 over thresholds from 0.05 to 0.94 in increments of 0.01. Table 46 reports test performance at the selected threshold for each run.

N Fixed-Checkpoint Study Protocol

Scope and selection. This targeted study was specified after the historical test results became available. The study focuses on Mamba-style, the leading head in the earlier movie comparison. The selected concatenation configuration comes from the completed six-configuration validation search. Default concatenation uses width 256 and learning rate 10^{-3}. Selected concatenation uses width 512 and learning rate 3×10^{-4}. The additive predictor has three independent width-256 branches and learning rate 10^{-3}. The six new Mamba-style fits use the original three-token bidirectional implementation. Each has width 256, state size 16, two blocks per direction and learning rate 10^{-3}. All models use binary cross-entropy, AdamW with weight decay 10^{-4}, batch size 256 and 30 epochs. Validation macro-AUC selects each checkpoint, retaining the earliest exact tie. The resulting comparison uses a fixed training recipe with model-specific capacities and the stated selection budgets.

Identity and precision. Eighteen baseline checkpoints are reused after their training-bank, validation-bank and weight hashes are verified. The additional-baseline runner initialises layers in a different random-number order from the historical shared-recipe implementation. Each Mamba-style head undergoes one training run and validation checkpoint selection. The checkpoint manifest records the 24 selected model identities. Every partition reuses the same weights and direct original logits, verified for equality across records. Model execution uses float32 with deterministic algorithms enabled and TF32 disabled. Projection sums accumulate in float64, and AUC is computed directly from logits. Logit scoring avoids float32 sigmoid saturation ties, allowing small differences from historical probability-based scores.

Reused baselines retain their original float32 sigmoid-based validation selections, and new Mamba-style fits use validation logits. All new projection contrasts use saved logits with these checkpoint selections fixed. Within each bank, independent aggregation checks validate the checkpoint seal, test-bank hashes and prediction hashes before estimating its intervals. These checks also verify original-logit equality across partitions and the deterministic reference indices.

Reference subsets. Let R be one selected reference subset of the test pool. For a scored example i, the projected logit is

$$P_{p,R}f(x_i) = \frac{1}{|R|}\sum_{j\in R} f(x_{i,U}, x_{j,V}) + \frac{1}{|R|}\sum_{j\in R} f(x_{j,U}, x_{i,V}) - \frac{1}{|R|^2}\sum_{j,k\in R} f(x_{j,U}, x_{k,V}). \quad (11)$$

Full-grid accumulation also collects these sums for reference sizes 256 and 1,024, using draw seeds 7101, 7102 and 7103. Shared reference indices keep each model comparison paired throughout subset evaluation. These subsets overlap with the scoring set and vary the empirical reference within the same test pool. All reference choices use the same fitted predictors and their cross-pair evaluations.

Conditional intervals. We draw 1,000 movie-row multiplicity vectors with bootstrap seed 20260910. Within each bank, the same vector weights every original and projected prediction across all models and fitted seeds. Each draw recalculates AUC from the fixed projected logits using these row weights. The intervals therefore condition on the fitted models and empirical projection reference from the test pool. A tie-aware weighted-rank AUC implementation matches explicit scikit-learn resampling within 3×10^{-16} in numerical tests. A paired draw contributes when both outcome classes occur in each of the 23 genre columns. Within-draw contrasts are averaged over the three fitted seeds before percentile calculation. Each pointwise interval spans the 2.5th to 97.5th percentiles of these averaged contrasts. These intervals describe each contrast under row reweighting with fitted seeds and the full-test projection reference fixed. Reference-subset variation is summarised separately, and all prespecified contrasts are reported.

Numerical and execution verification.

All 72 prediction archives match their checkpoint and test-bank identities. Direct original logits are identical across the three partitions of each fitted model. Cached-grid diagonals agree with direct logits within 1.91×10^{-5}. Independent recomputation of archived AUCs differs by at most 3.33×10^{-16}. Strict loading into the original model classes also confirms the parameter counts.

Both banks retained all 1,000 bootstrap draws under the complete-label validity rule. The GPU was an NVIDIA RTX 5090 with 32 GB memory. The runtime used Python 3.13.5 and Torch 2.9.1+cu128. Each cross-pair chunk contained four row indices against all 7,799 columns. The six Mamba-style fits took 23.3 to 31.3 seconds each. Complete-grid accumulation took 5.2 to 8.8 seconds for default concatenation and 5.8 to 9.6 seconds for selected concatenation. Mamba-style required 322.2 to 388.7 seconds per grid. These timings cover grid accumulation, excluding compression and metric recomputation. A separate efficiency probe measured 7.24 GiB peak allocated memory for Mamba-style at this chunk size.

Table 46: Thresholded scores of the fixed-checkpoint models in Table 43. Values are mean ± sample SD across three fitted seeds. A global probability threshold maximises validation micro-F1 over 0.05 to 0.94 in steps of 0.01. This threshold is then used for both test F1 metrics.

Bank	Model	Micro-F1	Macro-F1
w2v	Default concat	0.5775 ± 0.0010	0.4510 ± 0.0115
w2v	Selected concat	0.5787 ± 0.0018	0.4540 ± 0.0081
w2v	Strictly additive	0.5801 ± 0.0022	0.4659 ± 0.0037
w2v	Mamba-style	0.5929 ± 0.0019	0.4850 ± 0.0173
ST	Default concat	0.6862 ± 0.0005	0.6280 ± 0.0043
ST	Selected concat	0.6898 ± 0.0005	0.6319 ± 0.0015
ST	Strictly additive	0.6850 ± 0.0011	0.6292 ± 0.0026
ST	Mamba-style	0.6910 ± 0.0016	0.6374 ± 0.0081

Table 47: Own-model projection sensitivities $D_p(f)$ in AUC percentage points. Entries are three-seed means with pointwise 95% paired row-reweighting intervals. Fitted models and the full-test reference remain fixed. Strictly additive logits are preserved analytically, giving zero sensitivity for all partitions.

Bank	Partition	Mamba-style	Default concat	Selected concat
w2v	$ab \mid c$	0.289 $[0.203, 0.374]$	0.199 $[0.158, 0.240]$	0.222 $[0.187, 0.258]$
w2v	$ac \mid b$	0.328 $[0.246, 0.408]$	0.239 $[0.168, 0.318]$	0.257 $[0.194, 0.325]$
w2v	$a \mid bc$	0.115 $[0.024, 0.199]$	0.114 $[0.041, 0.185]$	0.110 $[0.050, 0.172]$
ST	$ab \mid c$	0.181 $[0.136, 0.223]$	0.071 $[0.053, 0.090]$	0.072 $[0.057, 0.087]$
ST	$ac \mid b$	0.156 $[0.110, 0.195]$	0.045 $[-0.0002, 0.089]$	0.090 $[0.048, 0.133]$
ST	$a \mid bc$	0.230 $[0.175, 0.283]$	0.088 $[0.041, 0.134]$	0.133 $[0.088, 0.180]$

Table 48: Reference-subset sensitivity of Mamba-style minus selected concat, in AUC percentage points (pp). Each range describes variation across three prespecified reference draws after averaging the three fitted seeds. The ranges summarise reference sensitivity.

Bank	Partition	Full G_p	$\lvert R\rvert = 256$ range	$\lvert R\rvert = 1,024$ range
w2v	$ab \mid c$	1.359	$[1.345, 1.373]$	$[1.352, 1.358]$
w2v	$ac \mid b$	1.355	$[1.345, 1.361]$	$[1.347, 1.355]$
w2v	$a \mid bc$	1.421	$[1.404, 1.432]$	$[1.412, 1.420]$
ST	$ab \mid c$	0.005	$[0.002, 0.004]$	$[0.002, 0.005]$
ST	$ac \mid b$	0.049	$[0.045, 0.052]$	$[0.047, 0.050]$
ST	$a \mid bc$	0.019	$[0.012, 0.019]$	$[0.016, 0.020]$

O TRAINED KNOWN-MECHANISM CONTROLS

Each of two independent input blocks contains one Rademacher signal coordinate and three irrelevant Rademacher coordinates. This construction gives 16 states per block and 256 equiprobable joint states. For signal coordinates $a, b \in \{-1, 1\}$, the additive mechanism has conditional label probability $\sigma(2a+2b)$. The pure-product mechanism uses $\mathbf{1}[ab > 0]$, equivalent to XOR up to label reversal. The mixed mechanism uses conditional label probability $\sigma(1.25a + 1.25b + 2.5ab)$. Each mechanism uses training seeds 0, 1 and 2.

The strictly additive MLP, concatenation MLP and bilinear-plus-additive control contain 386, 321 and 402 parameters. The additive model sums two $4 \to 32 \to 1$ ReLU branches, and concatenation uses an $8 \to 32 \to 1$ ReLU MLP. The bilinear control adds a learned 4×4 product to the additive branches. All use AdamW with learning rate 0.003, weight decay 0.0001 and batch size 256. Training minimises mean binary cross-entropy, with dropout disabled for every model. Models within each task and seed share observations, one fixed configuration and 100 training epochs. The greatest sampled-validation AUC selects the checkpoint, with ties retaining the earliest epoch. The analysis includes all 27 selected fits. CPU computation uses float64 precision and one thread throughout the study. Training, validation and secondary test sets contain 16, 8 and 128 repeats of each joint state, respectively. Their sample sizes are 4,096, 2,048 and 32,768, with independently sampled outcomes for nonbinary label probabilities. All support points appear during training, making this a controlled function-fitting evaluation.

Primary population AUC assigns each support point positive weight $p(y = 1 \mid x)$ and negative weight $1 - p(y = 1 \mid x)$. Tied scores receive half credit in the weighted AUC calculation. This evaluation integrates out label sampling, while training and checkpoint selection use sampled labels. An independent weighted-pair implementation verifies all 54 original/projected population scores within 1.5×10^{-15}. Exact EMAP evaluates every cross-pair in the complete 16×16 support. Independent sampled-test outcomes, full prediction precision, learning curves and check records are retained locally. The release supplies aggregate records and code for reproducing the generated experiments.

The strictly additive predictor reaches product-task AUC 0.5315 ± 0.0027, with zero own-model sensitivity D_p. Equal mixed-task population AUCs arise from the same ordering of probability groups under different fitted logits. These controls evaluate fitted prediction functions over independent blocks with fully observed joint support.

O.1 DIRECT COMPARISON WITH GROUP SHAP

This secondary analysis uses the original 27 selected synthetic checkpoints. The additional diagnostic protocol was fixed after their population-AUC results were available. It binds the original inputs by SHA-256 before computing the new SHAP arrays. Training, checkpoint selection and label laws follow Appendix O.

Let Q be the uniform law on the complete 16×16 support. For the raw-logit function f, define $\rho(f) = \|f - P_Q f\|_{L^2(Q)}$. We compute the four values of the two-group marginal coalition game by enumerating omitted-block states. The symmetric off-diagonal SHAP convention gives

$$\Phi_{12}(u, v) = \tfrac{1}{2}\{v_{u,v}(\{1, 2\}) - v_{u,v}(\{1\}) - v_{u,v}(\{2\}) + v_{u,v}(\varnothing)\}. \tag{12}$$

This specialises the SHAP interaction definition of Lundberg et al. (2018). Under this product law, $\Phi_{12} = (f - P_Q f)/2$ and its RMS equals $\rho(f)/2$. The coalition and projection implementations agree pointwise within 5.4×10^{-16}. This comparison uses the same marginal game and raw-logit scale for both quantities. The tables report ρ, and the supplied CSV files retain both RMS quantities.

Exact population AUC weights each support point by its known positive and negative label probabilities. Explicit pair sums agree with a separate weighted rank calculation and the archived metrics within 1.5×10^{-15}. Checkpoint reconstruction in NumPy reproduces the archived logits within 1.8×10^{-15}. All 677 numerical checks and 170 original-file hash checks pass. Machine-readable outputs retain every support prediction, coalition value, model result and paired contrast.

Table 49 summarises all nine model conditions. Table 50 also compares both nonlinear models with the strict additive model. Tables 51 and 52 report every seed.

Table 49: Functional and task diagnostics on all nine synthetic conditions. A, C and B denote the additive MLP, concatenation MLP and bilinear control. Residual RMS uses raw logits, and D_p uses AUC percentage points (pp). Values are three-seed means $\pm$ sample SD. Each diagnostic integrates the complete independent-block support. Additive and mixed task names describe their generating log odds.

Label mechanism	Model	Residual RMS	D_p (pp)
Additive	A	$< 10^{-12}$	0.000 ± 0.000
Additive	C	0.168 ± 0.064	-0.052 ± 0.090
Additive	B	0.346 ± 0.160	0.000 ± 0.000
Product	A	$< 10^{-12}$	0.000 ± 0.000
Product	C	0.691 ± 0.163	50.187 ± 0.186
Product	B	0.472 ± 0.277	49.792 ± 0.443
Mixed	A	$< 10^{-12}$	0.000 ± 0.000
Mixed	C	1.683 ± 0.706	21.897 ± 0.000
Mixed	B	1.592 ± 0.496	21.897 ± 0.000

Table 50: All three paired model comparisons per mechanism. G, G_p and E_p use AUC percentage points. Values are three-seed means $\pm$ sample SD. Each comparison pairs the same training and data seed. Model codes follow Table 49.

Mechanism	Head / baseline	G	G_p	E_p
Additive	C/A	-0.004 ± 0.150	0.048 ± 0.083	-0.052 ± 0.090
Additive	B/A	0.048 ± 0.083	0.048 ± 0.083	0.000 ± 0.000
Additive	B/C	0.052 ± 0.090	0.000 ± 0.000	0.052 ± 0.090
Product	C/A	46.849 ± 0.269	-3.339 ± 0.270	50.187 ± 0.186
Product	B/A	46.849 ± 0.269	-2.944 ± 0.426	49.792 ± 0.443
Product	B/C	0.000 ± 0.000	0.395 ± 0.596	-0.395 ± 0.596
Mixed	C/A	21.781 ± 0.117	-0.116 ± 0.117	21.897 ± 0.000
Mixed	B/A	21.781 ± 0.117	-0.116 ± 0.117	21.897 ± 0.000
Mixed	B/C	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

Table 51: All 27 fixed synthetic models. Residual RMS ρ uses raw logits. $S(f)$, $S(P_Q f)$ and D_p use AUC percentage points. Values displayed as zero are rounded for presentation. Full-precision records are supplied. Model codes follow Table 49.

Mechanism	Model	Seed	ρ	$S(f)$	$S(P_Q f)$	D_p
Additive	A	0	0.000	86.151	86.151	0.000
Additive	A	1	0.000	86.007	86.007	0.000
Additive	A	2	0.000	86.151	86.151	0.000
Additive	C	0	0.236	86.151	86.151	0.000
Additive	C	1	0.159	86.151	86.151	0.000
Additive	C	2	0.108	85.995	86.151	-0.156
Additive	B	0	0.196	86.151	86.151	0.000
Additive	B	1	0.514	86.151	86.151	0.000
Additive	B	2	0.328	86.151	86.151	0.000
Product	A	0	0.000	52.911	52.911	0.000
Product	A	1	0.000	53.101	53.101	0.000
Product	A	2	0.000	53.442	53.442	0.000
Product	C	0	0.636	100.000	49.628	50.372
Product	C	1	0.874	100.000	50.000	50.000
Product	C	2	0.564	100.000	49.811	50.189
Product	B	0	0.211	100.000	50.366	49.634
Product	B	1	0.764	100.000	49.707	50.293
Product	B	2	0.442	100.000	50.549	49.451
Mixed	A	0	0.000	69.336	69.336	0.000
Mixed	A	1	0.000	69.213	69.213	0.000
Mixed	A	2	0.000	69.101	69.101	0.000
Mixed	C	0	2.385	90.998	69.101	21.897
Mixed	C	1	0.972	90.998	69.101	21.897
Mixed	C	2	1.692	90.998	69.101	21.897
Mixed	B	0	1.465	90.998	69.101	21.897
Mixed	B	1	2.139	90.998	69.101	21.897
Mixed	B	2	1.171	90.998	69.101	21.897

Table 52: All 27 within-seed comparisons in AUC percentage points. The head and baseline use the same population, reference and metric. Model codes follow Table 49.

Mechanism	Head / baseline	Seed	G	G_p	E_p	D_h	D_c
Additive	C/A	0	0.000	0.000	0.000	0.000	0.000
Additive	C/A	1	0.144	0.144	0.000	0.000	0.000
Additive	C/A	2	-0.156	0.000	-0.156	-0.156	0.000
Additive	B/A	0	0.000	0.000	0.000	0.000	0.000
Additive	B/A	1	0.144	0.144	0.000	0.000	0.000
Additive	B/A	2	0.000	0.000	0.000	0.000	0.000
Additive	B/C	0	0.000	0.000	0.000	0.000	0.000
Additive	B/C	1	0.000	0.000	0.000	0.000	0.000
Additive	B/C	2	0.156	0.000	0.156	0.000	-0.156
Product	C/A	0	47.089	-3.284	50.372	50.372	0.000
Product	C/A	1	46.899	-3.101	50.000	50.000	0.000
Product	C/A	2	46.558	-3.632	50.189	50.189	0.000
Product	B/A	0	47.089	-2.545	49.634	49.634	0.000
Product	B/A	1	46.899	-3.394	50.293	50.293	0.000
Product	B/A	2	46.558	-2.893	49.451	49.451	0.000
Product	B/C	0	0.000	0.739	-0.739	49.634	50.372
Product	B/C	1	0.000	-0.293	0.293	50.293	50.000
Product	B/C	2	0.000	0.739	-0.739	49.451	50.189
Mixed	C/A	0	21.662	-0.235	21.897	21.897	0.000
Mixed	C/A	1	21.785	-0.112	21.897	21.897	0.000
Mixed	C/A	2	21.897	0.000	21.897	21.897	0.000
Mixed	B/A	0	21.662	-0.235	21.897	21.897	0.000
Mixed	B/A	1	21.785	-0.112	21.897	21.897	0.000
Mixed	B/A	2	21.897	0.000	21.897	21.897	0.000
Mixed	B/C	0	0.000	0.000	0.000	21.897	21.897
Mixed	B/C	1	0.000	0.000	0.000	21.897	21.897
Mixed	B/C	2	0.000	0.000	0.000	21.897	21.897

Exact analytic controls. Five analytic cases use independent $a, b \in \{-1, 1\}$, each pair having probability $1/4$. These declared functions are separate from the 27 trained models. For the scale cases, $Y = \mathbf{1}[a > 0]$, $c = a$ and $h = t(a + 0.5ab)$ with $t = 1$ or 10. For the shared-product case, $h = 2ab$, $c = ab$ and $Y = \mathbf{1}[ab > 0]$. The changed-label case keeps these functions and the input law fixed. It replaces the label law by $Y = \mathbf{1}[ab < 0]$. For the final case, $Y = \mathbf{1}[a > 0 \vee b > 0]$, $h = 100(a + b) + 99ab$ and $c = -ab$.

Table 53 separates score-scale effects, shared sensitivity and label-law effects. In the final case, $\rho(h) = 99$ and $\rho(c) = 1$. The AUCs satisfy $S(h) = S(P_Q h) = 1$, $S(c) = 5/6$ and $S(P_Q c) = 1/2$. Thus $E_p = -1/3$ accompanies the larger head residual. Across the two product label laws, ρ and SHAP RMS remain fixed. Both models' D_p change from $+1/2$ to $-1/2$. This construction illustrates how task sensitivity depends on the label law. Real-task transfer requires its own shift assumptions and evaluation.

Table 53: Exact analytic controls. ρ_h and ρ_c use the specified function-output units. The five rightmost columns use AUC percentage points. Values integrate four states exactly and have no seed SD.

Case	ρ_h	ρ_c	D_h	D_c	G	G_p	E_p
Scale $t = 1$	0.500	0.000	0.000	0.000	0.000	0.000	0.000
Scale $t = 10$	5.000	0.000	0.000	0.000	0.000	0.000	0.000
Shared product	2.000	1.000	50.000	50.000	0.000	0.000	0.000
Changed labels	2.000	1.000	-50.000	-50.000	0.000	0.000	0.000
Larger head residual	99.000	1.000	0.000	33.333	16.667	50.000	-33.333

P Historical Movie Comparisons

The historical shared-recipe comparison contains nine heads on the two standardised movie banks. Each uses binary cross-entropy, AdamW with weight decay 10^{-4}, batch size 256, learning rate 10^{-3} and 30 epochs. Runs use seeds 0, 1 and 2. Validation macro-AUC selects the reported epoch for each run. Test metrics are evaluated at validation improvements and included in the final best-epoch record. Mamba-style leads this comparison and motivates the later fixed-model study. The historical records therefore provide context for that choice.

Table 54: Historical feature-bank comparison for all nine movie heads. Values are mean macro-AUC $\pm$ sample SD across three seeds. The sentence-transformer bank replaces the text representation and its dimensionality.

Head	Raw w2v	Standardised w2v	Standardised ST
Shared-recipe concat	0.8372 ± 0.0003	0.8529 ± 0.0013	0.9316 ± 0.0001
PPA	0.8579 ± 0.0006	0.8660 ± 0.0011	0.9321 ± 0.0006
Bilinear	0.8528 ± 0.0021	0.8459 ± 0.0008	0.9295 ± 0.0006
Perceiver-style	0.8409 ± 0.0088	0.8606 ± 0.0001	0.9285 ± 0.0002
Mamba-style	0.8556 ± 0.0027	0.8710 ± 0.0008	0.9344 ± 0.0007
TRACE	0.8495 ± 0.0023	0.8639 ± 0.0011	0.9313 ± 0.0007
CGFM	0.8502 ± 0.0011	0.8456 ± 0.0021	0.9261 ± 0.0000
CGFM-CA	0.8490 ± 0.0017	0.8634 ± 0.0013	0.9308 ± 0.0008
CA+projection	0.8570 ± 0.0022	0.8651 ± 0.0020	0.9326 ± 0.0008

Standardising w2v features raises shared-recipe concatenation AUC from 0.8372 to 0.8529. Bilinear and CGFM move from above shared-recipe concatenation to below it in all three seeds (Table 55). With ST features, the nine-head AUC range contracts from 0.0254 to 0.0083. The supplied representations consequently shape the scale and ordering of the fusion comparison.

Independent width-256 linear/ReLU transformations of the poster block or of every block precede a joint nonlinear MLP in the processing controls. Poster transformation adds 0.30 AUC pp on w2v, while all-block transformation adds 0.68 pp. The all-block control has $ac|b$ sensitivity 0.18 pp.

Table 55: Effect of feature standardisation on w2v. Scores are mean ± sample SD across three seeds. Gaps compare each head with shared-recipe concatenation in its corresponding feature bank. Absolute scores use the 0–1 AUC scale. Gaps use AUC percentage points (pp).

Head	Unstandardised AUC	Standardised AUC	Unstandardised G (pp)	Standardised G (pp)
Shared-recipe concat	0.8372 ± 0.0003	0.8529 ± 0.0013	$+0.00$	$+0.00$
Bilinear	0.8528 ± 0.0021	0.8459 ± 0.0008	$+1.56$	-0.70
CGFM	0.8502 ± 0.0011	0.8456 ± 0.0021	$+1.30$	-0.73

With ST, poster transformation raises shared-recipe concatenation from 0.9316 to 0.9332, close to Mamba-style's 0.9344 (Table 56).

Table 56: Per-block processing and reference heads on the two standardised movie banks. Values are mean macro-AUC ± sample SD across three seeds. Independent block transformations feed a joint nonlinear MLP.

Head	MM-IMDb w2v	MM-IMDb ST
Shared-recipe concat	0.8529 ± 0.0013	0.9316 ± 0.0001
Project image + concatenate	0.8559 ± 0.0010	0.9332 ± 0.0006
Project all + concatenate	0.8597 ± 0.0012	0.9326 ± 0.0001
PPA	0.8660 ± 0.0011	0.9321 ± 0.0006
Mamba-style	0.8710 ± 0.0008	0.9344 ± 0.0007

Strictly additive and validation-selected baselines. Table 57 records one configuration per shared-recipe or additive baseline and six validation candidates for selected concatenation, each with three seeds.

Table 57: Baseline scores, capacity, and selection budget. Scores are mean macro-AUC ± sample SD across three seeds. Configuration counts give the validation-search budget per bank, with three training runs per configuration.

Bank	Baseline	Configs	Parameters	Macro-AUC
w2v	Shared-recipe concat	1	356,375	0.8529 ± 0.0013
w2v	Strictly additive	1	368,709	0.8573 ± 0.0008
w2v	Selected concat	6	712,727	0.8567 ± 0.0006
ST	Shared-recipe concat	1	476,183	0.9316 ± 0.0001
ST	Strictly additive	1	488,517	0.9323 ± 0.0004
ST	Selected concat	6	952,343	0.9332 ± 0.0002

Table 58: Historical w2v own-model projection sensitivity, in AUC percentage points (pp). Entries are mean ± sample SD across three seeds. A dash identifies a partition outside the recorded comparison.

Head	$D_{ab\|c}$	$D_{ac\|b}$	$D_{a\|bc}$
Shared-recipe concat	$+0.20 \pm 0.02$	$+0.19 \pm 0.06$	$+0.06 \pm 0.05$
PPA	$+0.31 \pm 0.03$	$+0.18 \pm 0.03$	$+0.09 \pm 0.01$
Bilinear	$+0.26 \pm 0.09$	-0.20 ± 0.08	-0.50 ± 0.15
Perceiver-style	$+0.33 \pm 0.09$	$+0.14 \pm 0.06$	–
Mamba-style	$+0.29 \pm 0.07$	$+0.33 \pm 0.08$	$+0.12 \pm 0.13$
TRACE	$+0.22 \pm 0.03$	$+0.18 \pm 0.06$	–
CGFM	$+0.51 \pm 0.06$	-0.35 ± 0.32	-0.54 ± 0.19
CGFM-CA	$+0.28 \pm 0.08$	$+0.02 \pm 0.11$	–
CA+projection	$+0.29 \pm 0.02$	$+0.15 \pm 0.07$	$+0.05 \pm 0.05$

Table 59: Historical ST own-model projection sensitivity, in AUC percentage points (pp). Entries are mean ± sample SD across three seeds. A dash identifies a partition outside the recorded comparison.

Head	$D_{ab\|c}$	$D_{ac\|b}$	$D_{a\|bc}$
Shared-recipe concat	$+0.08 \pm 0.01$	$+0.04 \pm 0.03$	$+0.09 \pm 0.03$
PPA	$+0.12 \pm 0.01$	$+0.09 \pm 0.02$	$+0.18 \pm 0.03$
Bilinear	$+0.39 \pm 0.03$	$+0.16 \pm 0.03$	$+0.38 \pm 0.04$
Perceiver-style	$+0.15 \pm 0.04$	$+0.04 \pm 0.03$	–
Mamba-style	$+0.18 \pm 0.02$	$+0.16 \pm 0.04$	$+0.23 \pm 0.04$
TRACE	$+0.09 \pm 0.04$	$+0.08 \pm 0.01$	–
CGFM	$+0.20 \pm 0.03$	-0.21 ± 0.05	-0.05 ± 0.07
CGFM-CA	$+0.22 \pm 0.02$	$+0.02 \pm 0.04$	–
CA+projection	$+0.13 \pm 0.01$	$+0.13 \pm 0.01$	$+0.21 \pm 0.01$

Q Historical Movie Projection Verification

Each movie grid contains $7,799^2 = 60,824,401$ cross-pairs. The historical runner accumulates row means, column means and diagonal scores in batches. Independent block encoders can be cached, while operations coupling blocks remain in the grid evaluation. Historical projection records come from matched training replays. The later fixed-model protocol reuses a single fitted checkpoint across partitions.

The movie archive contains 180 paired projection records: 72 for each bank's head comparison and 36 processing-control records. There are 108 recorded internal checks and 180 independent comparisons with the corresponding archived original scores. All satisfy the raw-AUC tolerance of 5×10^{-4}. Original AUC agrees across recorded partitions for all 66 movie head/seed combinations at four-decimal precision. CGFM-CA, Perceiver-style and TRACE have records for $ab|c$ and $ac|b$. The other six heads and the processing controls include all three partitions.

The archived 0–1 AUC scores retain four decimal places. Recomputing sensitivities from those four scores preserves the gap identity. Rounding can contribute up to 0.01 AUC pp error to a two-score contrast and 0.02 pp to excess sensitivity. Contrasts at this scale should therefore be interpreted at the precision of the archived records. The new fixed-model and prospective experiments retain unrounded scores and check the identity before display formatting.

R Additional Diagnostic Checks

Bisection and fully additive projections. Let a, b, c be independent and uniform on $\{-1, 1\}$, let $f(a, b, c) = ab + ac + bc$, and set $y = \mathbf{1}\{a = b = c\}$. The original score has AUC 1. Under the same uniform product reference, the three bisection projections are $P_{ab|c}f = ab$, $P_{ac|b}f = ac$, and $P_{a|bc}f = bc$. Each attains AUC $5/6$, with every positive example receiving score 1. Four of the

six negative examples receive score -1, and two receive score 1. The orthogonal projection onto functions of the form $u(a) + v(b) + w(c)$ is

$$P_{\text{add}} f = \mathbb{E}[f \mid a] + \mathbb{E}[f \mid b] + \mathbb{E}[f \mid c] - 2\mathbb{E}[f] = 0,$$

and its AUC is $1/2$. Each bisection retains a different within-group interaction, allowing high AUC under three distinct projected predictors. Full three-block additivity removes all three pairwise terms in this example. The calculation compares orthogonal projections of the same specified function f.

Sensitivity to behaviour outside the observed support. Suppose $A = B$ is uniform on $\{-1, 0, 1\}$ and $Y = \mathbf{1}\{A = 1\}$. For $\lambda \geq 0$, define

$$f_\lambda(a, b) = a + \lambda \mathbf{1}\{a \neq b\}\big(\mathbf{1}\{a \leq 0\} + \mathbf{1}\{b \leq 0\}\big).$$

Every f_λ agrees with $f_0(a, b) = a$ on the observed support, giving logits $(-1, 0, 1)$ and AUC 1. EMAP under the uniform 3×3 product of the observed marginals has diagonal logits $(-1 + 10\lambda/9,\ 10\lambda/9,\ 1 + 4\lambda/9)$. At $\lambda = 4$, these are $(31/9, 40/9, 25/9)$, yielding AUC 0, whereas the projection of f_0 retains AUC 1. Thus identical predictions on observed inputs can coexist with different EMAP outcomes because the product reference evaluates unobserved cross-pairs. Repeated reference sampling probes variability within a reference distribution that retains these unobserved cross-pairs. This finite-support construction isolates the dependence of EMAP on model behaviour outside the observed joint support.

Nonlinearity after a shared representation. A length-one key sequence gives a softmax weight of one before attention dropout. At evaluation the attention sublayer consequently uses only value and output projections. Subsequent nonlinear operations can still couple the input blocks. For example, $f(a, b) = \mathrm{ReLU}(a + b)$ has mixed difference

$$f(1, 1) - f(1, -1) - f(-1, 1) + f(-1, -1) = 2.$$

This example illustrates the modelling capacity retained by a joint nonlinear readout. We compared the original PPA class with a value-only replacement using synthetic inputs and random weights in float64. Complete-head logits agreed within 3.0×10^{-16} in evaluation and 2.8×10^{-16} in training with attention-weight dropout preserved. Removing that dropout gave a maximum difference of 0.6305. Query and key task-loss gradients were zero, although optimiser weight decay can update those weights. This check establishes the equivalence of the implemented computation paths under these conditions.

Executable checks. The algebraic diagnostic tests verify the XOR comparison in Section 3.2, score sensitivity with a nonzero residual, projection idempotence, and zero residual row and column means. They also check the paired score identity and the mixed difference of a joint ReLU. Baseline tests verify branch additivity, original concatenation equivalence, parameter counts, metric thresholds, and validation-only configuration selection.