LAMF: LAYER AUGMENTED MULTIMODAL FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning frequently suffers from modality imbalance, which existing methods primarily study during model training. However, we find that modality imbalance may prevent final predictions from fully exploiting the class information already learned in modality representations. We observe that predictions from some intermediate layers correctly classify examples that no convex combination of the original unimodal and fusion head outputs can classify correctly. Nevertheless, we show that high confidence training fits attenuate training loss responses to these predictions under bounded score changes. These observations suggest that fully exploiting learned modality information requires considering both which predictions to fuse and how to estimate their combination. We propose Layer Augmented Multimodal Fusion (LAMF), which fits linear probes to intermediate representations of frozen encoders to complement the original classifier outputs. LAMF applies base models and their probes to samples excluded from their fitting and uses the resulting predictions to estimate a global combination of all prediction sources. The method preserves the base architecture and training objective, and inference uses only one full model and its probes. Extensive experiments demonstrate that LAMF outperforms current state-of-the-art methods on multiple tasks.

1 INTRODUCTION

Multimodal learning integrates audio, images, and text to exploit complementary information across modalities (Baltrušaitis et al., 2019). However, additional inputs do not always translate into better predictions: differences in fitting and generalization across modalities, together with modality competition during joint optimization, can impede modality learning (Wang et al., 2020; Huang et al., 2022). Existing methods improve modality training through gradient modulation (Peng et al., 2022; Wei & Hu, 2024), alternating learning or decoupled optimization (Hua et al., 2024; Wei et al., 2025), and modified objectives or geometry-aware optimization (Fan et al., 2023; Chen et al., 2026; Yu et al., 2026). These methods primarily focus on model training. An important question remains: does the existing prediction combination fully exploit the class information already learned in modality representations?

To address this question, we examine how prediction sources and their combination affect the final decision. First, recombining final outputs alone may be insufficient to exploit class information in intermediate representations. On CREMA-D and Kinetics-Sounds, we freeze the trained encoders, identify examples that cannot be correctly classified by any convex combination of the unimodal and fusion head outputs, and classify these examples using linear probes at different layers (Alain & Bengio, 2016). As shown in Figure 1(a), intermediate layer probes correctly predict a subset of these examples, with the fraction varying across modalities and layers. For these examples, intermediate layer predictions offer correction opportunities beyond convex combinations of the unimodal and fusion head outputs. Second, training performance may not reliably determine how these predictions should be combined. In Figure 1(b), a broad range of weights achieves near-optimal training accuracy but substantially different test accuracy. This motivates examining the complementary value of intermediate layer predictions beyond the data used to fit the model and learning their combination from examples excluded from fitting the corresponding predictors.

These observations lead to two related choices in multimodal fusion: **which predictions to combine, and how to learn their combination.** Expanding prediction sources can provide correction opportunities unavailable to convex combinations of final outputs, while an appropriate basis for

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

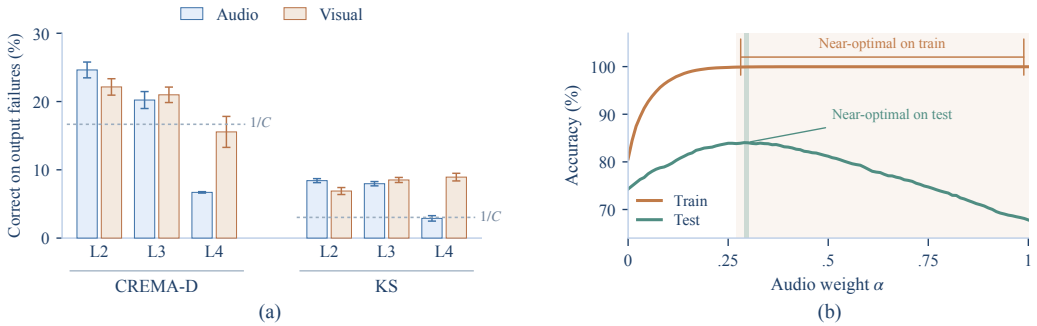


Figure 1: Intermediate layer complementarity and limits of in-sample combination estimation. (a) On CREMA-D and Kinetics-Sounds (KS), probe accuracy at different audio and visual layers is measured on examples that no convex combination of the unimodal and fusion head outputs classifies correctly. Dashed lines denote uniform guess accuracy $1/C$, with C classes. (b) Audio–visual fusion weights yield near-optimal training accuracy over a broad range but substantially different test accuracy.

estimation helps combine these sources. Our goal is to incorporate complementary intermediate layer predictions into the final decision and estimate their combination from performance beyond the fitting data, without changing the base architecture or training objective.

Based on these insights, we propose Layer Augmented Multimodal Fusion (LAMF), which complements final outputs with intermediate layer class predictions and estimates their combination from out-of-fold (OOF) performance. Specifically, LAMF fits linear predictors at prespecified layers of frozen encoders, mapping representations across modalities and layers into a shared class space. This allows class information decodable from intermediate representations to directly complement the original classifier predictions. To reduce the influence of strong in-sample fitting on combination estimation, we follow the data-separation principle of ensemble learning (Wolpert, 1992; van der Laan et al., 2007). Two-fold cross-prediction ensures that each example is predicted by a base model and its probes fitted without that example. Using these OOF records, LAMF places layer predictions, unimodal outputs, and the fusion head output in a common cross entropy objective and estimates a global combination from their combined performance beyond the fitting data.

Our contributions are threefold:

- We analyze multimodal information utilization through prediction sources and combination estimation, identifying how intermediate layer predictions complement existing classifier outputs. We further characterize loss response attenuation under high confidence training fits and conditions for improvement in deployment cross entropy.
- We propose LAMF, combining layer predictions from each modality with OOF combination estimation, and specify probe refitting and deployment on the full model.
- We conduct experiments on eight tasks spanning audio–visual, text–visual, and text–audio–visual learning, outperforming current state-of-the-art methods on multiple tasks and demonstrating improved utilization of modality information.

2 Related Work

Imbalanced multimodal learning. Research on modality imbalance primarily adjusts optimization and representation learning. OGM-GE modulates modality gradients (Peng et al., 2022), while MMPareto reconciles unimodal and multimodal gradients (Wei & Hu, 2024). ReconBoost alternates modality updates (Hua et al., 2024), and DGL decouples encoder and fusion optimization (Wei et al., 2025). MASAM adapts sharpness-aware perturbations to modality dominance (Chen et al., 2026). Representation-focused methods use class prototypes (PMR; Fan et al., 2023) or maximize total correlation among multimodal features and labels (TCMax; Yu et al., 2026). Our method instead examines how class information in learned representations across modalities and network

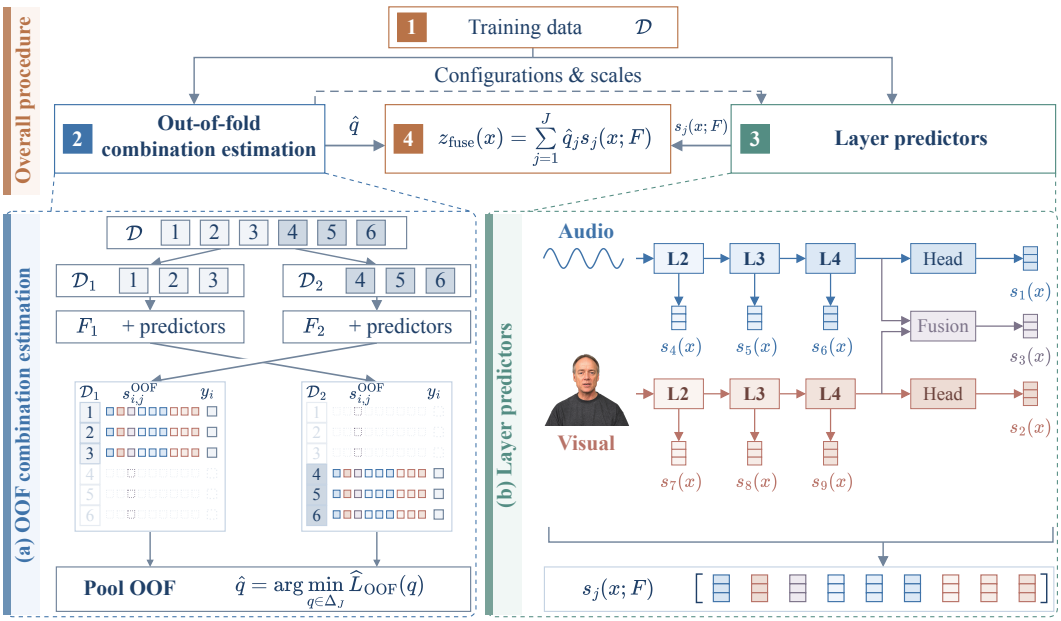


Figure 2: **LAMF** estimation and deployment (convex example).

depths contributes to the final prediction. It preserves the base architecture and training objective while expanding the information available for fusion through intermediate layer predictions and their combination, complementing training stage improvements.

Intermediate layer predictions and fusion. Linear probes use independently trained classifiers to analyze class information decodable at different network layers (Alain & Bengio, 2016). Head2Toe selects features across layers of pretrained networks for target-task classification (Evci et al., 2022). LATES fits intermediate layer probes on the original training set and learns nonnegative weights on a held-out set, combining probe and final classifier logits to improve image classification accuracy and calibration (Khalifa et al., 2022); our comparison is in Appendix C.2. For combination estimation, stacked generalization, stacked regression, and Super Learner train a higher-level combination using base learner predictions on examples excluded from fitting those learners, establishing a data-separation principle for combination estimation (Wolpert, 1992; Breiman, 1996; van der Laan et al., 2007). LAMF builds on these ideas to study which modality and layer predictions complement existing heads and how estimation records affect their gains. We connect these gains to training loss responses and estimation and deployment errors, and test the two components through controlled comparisons of prediction sources, estimation records, and deployment procedures.

3 Method

The preceding observations show that final outputs can leave complementary information in modality representations unused. Moreover, combinations with similar training accuracy can differ substantially in test accuracy. To address these limitations, LAMF supplements the original classifier outputs with intermediate layer predictions and estimates their combination from OOF records. The overall framework is illustrated in Fig. 2.

3.1 Problem Formulation

We use convex aggregation for analysis; the framework extends to regularized linear stacking. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ contain N samples with M modalities and C classes. Each input $x_i = (x_{i,1}, \ldots, x_{i,M})$ has modality components $x_{i,m}$ and label $y_i \in \{1, \ldots, C\}$. Each of J prediction sources s_j maps inputs to scores $s_j(x) \in \mathbb{R}^C$ in a shared class space. For bimodal tasks, sources comprise the existing unimodal and fusion heads and linear probes at prespecified modality–layer

locations, indexed consistently across fold and full models. The convex implementation combines their scores using global fusion weights $q = (q_1, \ldots, q_J)$, where q_j weights source j:

$$z_q(x) = \sum_{j=1}^{J} q_j s_j(x), \qquad q \in \Delta_J = \left\{ q \in \mathbb{R}^J : q_j \geq 0, \sum_{j=1}^{J} q_j = 1 \right\}. \tag{1}$$

The weights are shared across samples. We learn them using cross entropy $\ell(z, y) = \log \sum_{c=1}^{C} \exp(z_c) - z_y$, where z_c is the score for class c. These score combination coefficients also depend on score scales and correlations between sources.

3.2 Constructing Intermediate Layer Predictions

LAMF constructs intermediate layer predictors, or layer predictors, on fixed encoders to access class information decodable beyond the existing heads. For an intermediate layer source j, let $h_j(x) \in \mathbb{R}^{d_j}$ be the pooled representation at its modality–layer location, with dimension d_j; source index j identifies this modality–layer pair throughout. Let $\mathcal{D}_{\text{fit}}$ contain the n samples used to train the corresponding base model. Using only these samples, estimate standardization statistics to obtain $\widetilde{h}_j(x)$ and fit a ridge probe:

$$(\widehat{W}_j, \widehat{b}_j) = \arg\min_{W,b} \left\{ \frac{1}{n} \sum_{(x_i, y_i) \in \mathcal{D}_{\text{fit}}} \left\| e_{y_i} - W^\top \widetilde{h}_j(x_i) - b \right\|_2^2 + \tau \|W\|_F^2 \right\}. \tag{2}$$

Here, $e_y \in \mathbb{R}^C$ is 1 at class y and zero elsewhere; $W \in \mathbb{R}^{d_j \times C}$, $b \in \mathbb{R}^C$, and $\tau > 0$ is the regularization strength for the mean squared loss. The intercept is unpenalized; $\top$ denotes transpose, and the squared norms sum squared entries. Probes produce raw scores $r_j(x) = \widehat{W}_j^\top \widetilde{h}_j(x) + \widehat{b}_j$. Fusion scales them as $s_j(x) = \widehat{\gamma}_j r_j(x)$, with positive $\widehat{\gamma}_j$ estimated in Section 3.3. Standardization statistics and probe parameters are determined only from $\mathcal{D}_{\text{fit}}$ and applied unchanged to samples excluded from the corresponding base model's fitting.

To assess this probe's contribution, fix the model, probes, and scales. Let $z_0(x)$ combine the existing heads convexly, with $p_0(x) = \text{softmax}(z_0(x))$. The probe supplies the score correction $a_j(x) = s_j(x) - z_0(x)$. Assigning it weight η gives

$$z_{\eta,j}(x) = (1 - \eta) z_0(x) + \eta s_j(x) = z_0(x) + \eta a_j(x), \qquad 0 \leq \eta \leq 1. \tag{3}$$

Proposition 1 (Local contribution of an intermediate layer prediction). Let P be the distribution of a new sample (X, Y). Define $L_j(\eta) = \mathbb{E}_P \ell(z_{\eta,j}(X), Y)$ and $\text{osc}(a) = \max_c a_c - \min_c a_c$ for $a \in \mathbb{R}^C$. If $L_j(0) < \infty$ and $\mathbb{E}_P \, \text{osc}(a_j(X)) < \infty$, the initial rate of decrease in expected cross entropy when increasing this source's weight is

$$A(s_j; z_0) = -L_j'(0^+) = \mathbb{E}_P \left[(e_Y - p_0(X))^\top a_j(X) \right]. \tag{4}$$

Here, $L_j'(0^+)$ denotes the right derivative at zero. If $A(s_j; z_0) > 0$, sufficiently small $\eta > 0$ gives $L_j(\eta) < L_j(0)$; if $A(s_j; z_0) \leq 0$, no decrease in expected cross entropy occurs along the fixed interpolation in Equation (3). The proof is in Appendix A.1.

Equation (4) weights the probe correction a_j by the current combination's class residual. Increasing the true class score relative to high-probability incorrect classes contributes positively to A. Thus, A measures a source's local contribution to expected cross entropy, rather than its standalone accuracy or independent class information. This contribution can include confidence calibration and does not alone guarantee an accuracy gain. Appendix E examines realized corrections and calibration controls to distinguish classification gains from changes in confidence.

3.3 Out-of-Fold Combination Estimation

Intermediate layer predictions extract modality information useful for classification, but whether it improves the final prediction also depends on how their combination is estimated. For convex aggregation, we examine the loss change induced by increasing the same probe's weight when weights are adjusted using training cross entropy. Fix the model, probes, and scales fitted on $\mathcal{D}$, and define

$\widehat{L}_{\mathrm{tr},j}(\eta) = N^{-1}\sum_{i=1}^{N}\ell(z_{\eta,j}(x_i), y_i)$ as the mean training cross entropy along the same path. Its initial rate of decrease is

$$\widehat{A}_{\mathrm{tr},j} = -\widehat{L}'_{\mathrm{tr},j}(0^+) = \frac{1}{N}\sum_{i=1}^{N}(e_{y_i} - p_0(x_i))^\top a_j(x_i). \tag{5}$$

A positive $\widehat{A}_{\mathrm{tr},j}$ means that a small increase in this source's weight lowers training cross entropy; a magnitude near zero indicates a small initial loss change under this adjustment.

Proposition 2 (Attenuation of the initial decrease rate under high confidence training fits). If $\mathrm{osc}(a_j(x_i)) \le B_j < \infty$ for every training sample, with $B_j \ge 0$, then

$$\begin{aligned} |\widehat{A}_{\mathrm{tr},j}| &\le \frac{B_j}{N}\sum_{i=1}^{N}(1 - p_{0,y_i}(x_i)), \\ 0 \le \widehat{L}''_{\mathrm{tr},j}(0) &\le \frac{B_j^2}{N}\sum_{i=1}^{N}(1 - p_{0,y_i}(x_i)). \end{aligned} \tag{6}$$

Here, $p_{0,y_i}(x_i)$ is the true class probability. If these probabilities approach 1 while the same B_j is maintained, both the initial rate of decrease and curvature vanish. A sufficiently fitted combination of original heads, including the fusion head, can reach this regime. The small residuals attenuate training loss responses to bounded intermediate layer corrections, but do not determine which source receives greater weight. LAMF therefore estimates combinations using samples excluded from model and probe fitting. Equation (6) follows by averaging samplewise derivative bounds; Appendix A.2 gives the finite-change analysis for general weight adjustments.

Two-fold cross-prediction. Following the data-separation principle of stacked generalization and Super Learner (Wolpert, 1992; van der Laan et al., 2007), split $\mathcal{D}$ within each class into complementary halves $\mathcal{D}_1, \mathcal{D}_2$. Train a copy of the base model on each half using the original architecture and training procedure, termed fold models F_1, F_2, and then freeze their encoders. For $k \in \{1, 2\}$, estimate standardization statistics on $\mathcal{D}_k$ and fit each candidate probe by Equation (2) with $\mathcal{D}_{\mathrm{fit}} = \mathcal{D}_k$. For representations and probe parameters, the second subscript k identifies the fold model. For $(x_i, y_i) \in \mathcal{D}_{3-k}$, probe j's raw OOF score is

$$r_{i,j}^{\mathrm{OOF}} = \widehat{W}_{j,k}^\top \widetilde{h}_{j,k}(x_i) + \widehat{b}_{j,k}. \tag{7}$$

The existing heads also predict the opposite half. For each source and candidate configuration, merge both blocks in the original sample order and pair them with labels. Each raw score is generated without fitting the corresponding model or probe on that sample.

Score scales and fusion weights. Select each probe's pooling and regularization configuration using pooled OOF records and labels. Both folds use the same selected configuration but retain their own fitted parameters. Fit $\widehat{\gamma}_j$ by minimizing the selected probe's mean OOF cross entropy over $[\gamma_{\min}, \gamma_{\max}]$, where $0 < \gamma_{\min} \le \gamma_{\max} < \infty$. Set $s_{i,j}^{\mathrm{OOF}} = \widehat{\gamma}_j r_{i,j}^{\mathrm{OOF}}$; existing heads retain unit scale. With configurations and scales fixed, jointly optimize all sources:

$$\widehat{L}_{\mathrm{OOF}}(q) = \frac{1}{N}\sum_{i=1}^{N}\ell\left(\sum_{j=1}^{J} q_j s_{i,j}^{\mathrm{OOF}}, y_i\right), \qquad q \in \Delta_J. \tag{8}$$

This convex objective is solved for $\widehat{q} \in \Delta_J$ satisfying $\widehat{L}_{\mathrm{OOF}}(\widehat{q}) \le \inf_{q\in\Delta_J}\widehat{L}_{\mathrm{OOF}}(q) + \xi$, where $\xi \ge 0$ bounds optimization error. Along Equation (3)'s interpolation, Equation (4)'s OOF counterpart is the objective's negative directional derivative. All sources are optimized jointly without screening by A; configuration, scale, and weight estimation use OOF labels. Under stronger assumptions, OOF estimation is risk-consistent within the selected combination family (Appendix A.5).

3.4 Model Refitting and Inference

Train the full model F on all of $\mathcal{D}$ using the base procedure; freeze its encoders and refit the probes, including their standardization statistics, coefficients, and intercepts. Retain the prespecified layers

and OOF-selected pooling, regularization, scales $\widehat{\gamma}_j$, and weights $\widehat{q}$. Denote the existing head outputs and scaled probe predictions by $s_j(x; F)$. Inference uses only F and its probes:

$$z_{\text{fuse}}(x) = \sum_{j=1}^{J} \widehat{q}_j s_j(x; F), \qquad \widehat{y}(x) = \arg\max_{1 \le c \le C} [z_{\text{fuse}}(x)]_c. \tag{9}$$

To analyze how supplementary modality information translates into deployment gains, fix all models, probes, configurations, and scales. Let $L(q)$ be the full model's expected cross entropy on new samples from P, combining all sources with weights $q \in \Delta_J$. Let $L_+^\star$ and $L_0^\star$ be its minimum over global convex combinations using all sources and only existing heads, respectively. The ideal gain is $G = L_0^\star - L_+^\star \ge 0$. Under Proposition 1's conditions, if a full model head combination z_0 attains $L_0^\star$ and a layer prediction satisfies $A(s_j; z_0) > 0$, then $G > 0$.

Let $\overline{L}(q)$ average the fold models' expected cross entropies on independent samples from the training distribution, combining all selected sources with the same q. The losses of F_1 and F_2 receive weights $|\mathcal{D}_2|/N$ and $|\mathcal{D}_1|/N$, respectively, matching their numbers of OOF records. We average fold losses rather than first averaging their logits. Define

$$\epsilon = \sup_{q \in \Delta_J} \left| \widehat{L}_{\text{OOF}}(q) - \overline{L}(q) \right|, \qquad \Gamma = \sup_{q \in \Delta_J} \left| \overline{L}(q) - L(q) \right|. \tag{10}$$

Equation (10) compares the empirical OOF loss with the expected fold loss through ϵ, and the expected fold loss with the expected full model loss through Γ. The latter includes changes from model and probe refitting and any training–test distribution difference.

Proposition 3 (Conditional risk comparison for the full model). If these expected losses and discrepancies are finite and $\widehat{q}$ satisfies Section 3.3's optimization condition, then

$$L(\widehat{q}) - L_0^\star \le 2\epsilon + 2\Gamma + \xi - G. \tag{11}$$

For the selected configurations and scales, $G > 2\epsilon + 2\Gamma + \xi$ suffices to outperform the best convex combination of existing heads in expected cross entropy (Appendix A.3). Under stronger assumptions, we derive a positive lower bound on LAMF's population accuracy gain over an output only baseline (Appendix A.5). Linear stacking instead deploys a regularized cross class linear map with its OOF-fitted score standardizer; the map is shared across samples and kept fixed at inference, while full model probe refitting remains unchanged (Appendix A.4).

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate LAMF on eight audio–visual (AV), text–visual (TV), and text–audio–visual (TAV) classification tasks (Table 1). AVSBench-CLS is used for clip classification, and Food-101 uses the UPMC image–text version. We report accuracy (%) on all tasks and macro-F1 in selected ablations; sentiment fusion results use Has0. Tables 2, 3 compare convex aggregation and linear stacking; other analyses use convex aggregation unless stated otherwise.

Backbones and inputs. Audio–visual tasks use separate audio and visual ResNet18 encoders (He et al., 2016), with log-magnitude short-time Fourier transform (STFT) spectrograms as audio inputs. CREMA-D uses a single frame and audio repeated or truncated to 20 seconds; the other audio–visual tasks use three frames. Food-101 uses ResNet18 and a bidirectional GRU (Cho et al., 2014) for images and text, respectively. These encoders and the Food-101 word embeddings are trained from scratch. For MVSA-Single, we fine-tune pretrained BERT-base-uncased (Devlin et al., 2019) and ImageNet-pretrained ResNet152. CMU-MOSI/MOSEI use three independent Transformer branches (Vaswani et al., 2017) trained from scratch on fixed, aligned MMSA features, including pre-extracted BERT text features.

Table 1: Statistics of the datasets.

Dataset	Task	Classes	Train	Val.	Test
CREMA-D	AV	6	6,698	–	744
KSounds	AV	33	20,853	2,318	1,633
AVSBench	AV	23	3,452	740	740
VGGSound50	AV	50	25,954	–	2,499
Food-101	TV	101	62,971	5,000	22,715
MVSA-Single	TV	3	1,539	518	519
CMU-MOSI	TAV	2	1,284	229	686
CMU-MOSEI	TAV	2	16,326	1,871	4,659

Table 2: Mean classification accuracy (%) on four audio–visual datasets. The best results and second best results are **bold** and underlined, respectively.

Methods	CREMA-D			Kinetics-Sounds			AVSBench			VGGSound50		
	Audio	Visual	**Multi**	Audio	Visual	**Multi**	Audio	Visual	**Multi**	Audio	Visual	**Multi**
Unimodal	67.20	**73.79**	–	51.01	52.17	–	76.98	52.84	–	42.59	36.04	–
Concat	–	–	68.55	–	–	61.30	–	–	83.51	–	–	50.47
OGM-GE (Peng et al., 2022)	53.94	35.08	76.66	49.09	31.80	69.30	74.46	25.27	88.15	40.12	21.29	54.81
MMPareto (Wei & Hu, 2024)	67.61	48.30	73.61	57.01	50.75	73.10	82.07	50.32	90.81	46.98	34.23	58.54
InfoReg (Huang et al., 2025)	69.00	55.69	77.64	55.99	51.42	73.85	**83.02**	52.97	91.71	46.33	36.53	59.90
DGL (Wei et al., 2025)	68.82	69.76	81.41	56.89	59.85	77.51	81.89	64.14	**94.14**	**48.13**	40.66	61.24
AUG (Jiang et al., 2025)	**69.27**	73.48	84.41	56.11	**64.48**	77.67	81.17	58.78	88.02	43.63	39.80	58.06
TCMax (Yu et al., 2026)	68.59	72.58	83.51	54.38	52.48	76.12	68.87	36.22	87.16	46.10	31.28	60.40
IBML (Qin et al., 2026)	65.86	66.22	80.06	53.38	59.89	77.44	78.69	59.64	92.34	43.74	39.83	60.54
CPSC (Jiang et al., 2026)	67.16	71.33	83.51	**58.77**	62.48	74.69	80.90	**67.43**	87.30	46.77	**42.52**	56.68
Ours$_{\text{convex}}$	68.33	73.34	85.66	55.34	57.60	77.74	80.38	54.00	92.39	44.74	38.91	60.61
Ours$_{\text{linear stacking}}$	68.23	73.52	**86.38**	55.34	57.40	**78.81**	79.77	53.11	92.21	44.60	38.00	**62.18**

Table 3: Mean accuracy (%) on text–visual and trimodal datasets; conventions follow Table 2.

Methods	UPMC-Food101			MVSA-Single			CMU-MOSI				CMU-MOSEI			
	Text	Image	**Multi**	Text	Image	**Multi**	Text	Audio	Visual	**Multi**	Text	Audio	Visual	**Multi**
Unimodal	85.66	56.21	–	76.24	63.71	–	78.47	58.50	55.44	–	**75.00**	**57.54**	59.83	–
Concat	–	–	87.73	–	–	71.55	–	–	–	78.57	–	–	–	75.18
OGM-GE (Peng et al., 2022)	81.26	35.36	87.87	74.95	59.79	78.74	77.55	47.33	48.15	78.67	74.53	50.97	57.30	75.08
MMPareto (Wei & Hu, 2024)	84.23	57.46	90.26	77.14	63.84	77.59	78.96	56.03	50.19	79.83	74.61	52.30	58.18	75.44
InfoReg (Huang et al., 2025)	84.39	51.09	90.29	76.94	61.14	80.22	78.81	55.83	53.30	79.64	74.40	55.31	59.33	74.85
DGL (Wei et al., 2025)	83.93	**57.92**	90.07	76.75	63.97	78.16	77.89	51.65	50.73	78.86	74.81	49.02	48.98	75.05
AUG (Jiang et al., 2025)	84.55	56.34	90.45	**77.25**	62.30	80.59	78.23	51.36	48.69	78.23	74.59	53.95	59.40	75.01
TCMax (Yu et al., 2026)	83.52	53.10	90.19	77.02	**64.22**	80.33	77.70	**59.43**	52.82	78.81	74.58	54.35	58.10	75.55
IBML (Qin et al., 2026)	84.02	55.73	90.67	75.21	59.54	78.36	78.62	55.30	**56.41**	79.25	74.18	51.50	59.35	74.89
CPSC (Jiang et al., 2026)	83.98	50.98	89.41	76.17	63.84	79.13	77.70	52.92	51.51	78.47	74.72	54.47	59.16	75.20
Ours$_{\text{convex}}$	85.71	56.14	**92.08**	76.23	63.90	79.58	**79.17**	50.20	48.12	**80.56**	74.74	56.80	**60.01**	75.90
Ours$_{\text{linear stacking}}$	**85.72**	56.28	91.24	74.82	62.04	**81.70**	75.80	48.20	48.01	80.23	72.71	54.79	59.08	**76.02**

4.2 Main Results

We compare LAMF with independently trained unimodal models, concatenation fusion, and eight methods addressing modality imbalance: OGM-GE (Peng et al., 2022), MMPareto (Wei & Hu, 2024), InfoReg (Huang et al., 2025), DGL (Wei et al., 2025), AUG (Jiang et al., 2025), TCMax (Yu et al., 2026), IBML (Qin et al., 2026), and CPSC (Jiang et al., 2026). Tables 2, 3 report individual branch and fusion accuracy across the eight classification tasks.

Among the compared methods, linear stacking LAMF achieves the highest mean fusion accuracy on five tasks; convex LAMF leads on Food-101 and MOSI. Linear stacking exceeds AUG on CREMA-D by 1.97 percentage points, and convex LAMF exceeds IBML on Food-101 by 1.41 points. Convex LAMF also surpasses AUG on CREMA-D despite lower accuracy in both original branches. Tables 2, 3 compare complete implementations, not isolated aggregator changes; Table 4 and Appendix C.1 fix base models to assess prediction sources and their combinations.

4.3 Utilization of Learned Modality Information

We freeze the encoders and classifiers of OGM-GE, DGL, and TCMax and fit each convex combination on the method's own OOF records (Table 4). The gains are Δ_{rw}: Output only minus Original, and Δ_{layer}: +Both minus Output only. Layer predictors add 0.54–1.08 accuracy points across all six models, even with negligible reweighting gains, supporting better use of learned modality information than the estimated output only combination. For TCMax, audio probes add 0.94 points on CREMA-D, versus 0.14 from visual probes; on KS, visual probes add 0.92 points while audio probes add none. Thus, the useful layer predictions vary across tasks and modalities. Using both probe sets gives the highest accuracy in five of the six frozen models.

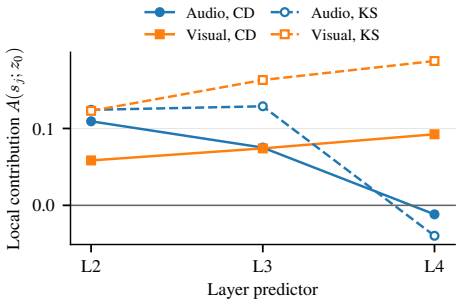
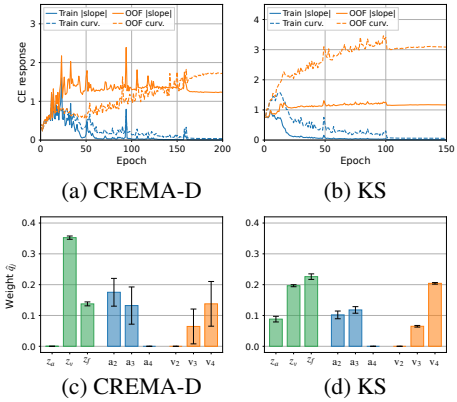


Figure 3: Local contribution $A(s_j; z_0)$ (Eq. 4), estimated on test samples and averaged over three methods. Positive values indicate an initial cross entropy decrease from Output only toward a layer prediction. CD: CREMA-D.



(a) CREMA-D (b) KS

(c) CREMA-D (d) KS

Figure 4: Fusion diagnostics: (a,b) In-sample and OOF loss responses; (c,d) OOF-estimated fusion weights (mean ± std).

Table 4: Frozen model accuracy (%). Original: native fusion; Output only: three combined heads. +A/+V probes add three audio/visual layer predictions; +Both adds both sets. Gains in pp.

Dataset	Model	Original	Output only	+A probes	+V probes	**+Both**	Δ_{rw}	Δ_{layer}
CREMA-D	OGM-GE	76.88	76.88	77.42	77.28	**77.96**	+0.00	+1.08
	DGL	81.18	84.14	84.27	84.14	**84.68**	+2.96	+0.54
	TCMax	83.20	83.33	84.27	83.47	**84.41**	+0.13	+1.08
KS	OGM-GE	68.77	68.83	69.26	69.14	**69.44**	+0.06	+0.61
	DGL	76.79	77.59	78.32	77.59	**78.63**	+0.80	+1.04
	TCMax	75.38	75.38	75.38	**76.30**	76.18	+0.00	+0.80

Figure 3 estimates test set $A(s_j; z_0)$ (Equation 4) relative to Output only; positive values indicate local improvement, not accuracy gains. Averaged over three methods, audio L2/L3 contributions are positive and L4 negative on both datasets, while visual contributions increase with depth. Figure 1(a) shows correction opportunities; Appendix E.1 quantifies actual net corrections; E.3 adds independent evaluations and analyses on Food-101 and MOSI.

4.4 Fusion Response and Weight Allocation

Figure 4(a,b) tracks mean absolute samplewise slopes and mean curvature of cross entropy toward the audio output from the audio–visual combination, each epoch on in-sample and OOF records. In-sample responses decline late in training; OOF responses remain larger. This decline is consistent with Proposition 2: under bounded perturbations, high confidence fitting attenuates training loss responses to layer predictions (Appendix E.2). Figure 4(c,d) shows OOF weights retained for full model inference (a/v: audio/visual; three heads and six layer predictors). Audio L4 and visual L2 receive negligible weights on both tasks, qualitatively consistent with their lower local contributions within each modality in Figure 3. These weights describe the combination, not independent information contributions; Appendix E.4 evaluates probe refitting and combination transfer.

4.5 Ablation Studies

In Table 5, w/o OOF fits weights in sample with fixed probe configurations and scales; Appendix D.1 checks selection reuse and solver convergence. In Table 5, adding both probe sets improves both OOF aggregators but leaves fold in-sample accuracy unchanged. With fold models, sources, configurations, and scales fixed, switching convex estimation to OOF records increases independent-evaluation layer gains from 0.23 to 1.70 points on CREMA-D and from 0.00 to 1.50 on KS (Appendix E.2). Nine source convex fusion loses accuracy without the fusion head output or scale fitting, or with uniform weights. The fusion head ablation supports complementing, rather than replacing, the model's original joint prediction. For fixed sources, positive scaling preserves the classification

Table 5: Fusion ablations (accuracy, %). Output only combines head outputs; +A/+V probes add three audio/visual layer predictions; +Both adds both sets. Without the fusion head: 2/5/5/8 sources.

Variant	CREMA-D				KS			
	Output only	+A probes	+V probes	+Both	Output only	+A probes	+V probes	+Both
w/o OOF (full)	71.42	71.42	71.42	71.42	69.24	73.42	73.42	73.42
w/o OOF (fold)	71.19	71.19	71.19	71.19	73.40	73.40	73.40	73.40
w/o fusion head	84.41	85.26	84.05	85.04	75.83	75.65	75.79	76.06
w/o scale fitting	84.41	84.77	84.59	84.77	76.57	76.53	76.59	76.59
w/o learned weights	76.75	75.85	80.91	80.20	75.14	72.34	76.42	75.34
LAMF$_{\text{convex}}$	84.41	85.44	84.23	**85.66**	76.57	76.67	76.67	**77.74**
LAMF$_{\text{linear stacking}}$	84.77	85.89	85.08	**86.38**	76.83	77.42	78.12	**78.81**

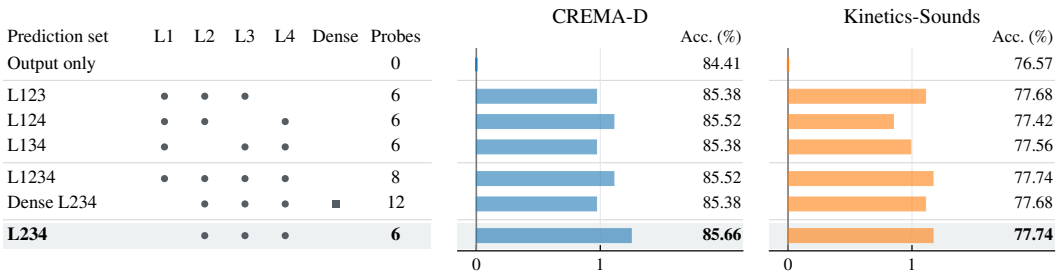


Figure 5: Prediction sets and accuracy differences (pp) from Output only. Dots mark stages; right columns report accuracy (%).

decisions attainable by simplex combinations, but changes the cross entropy objective and hence combination estimation (Appendix D.2). Ridge performs best among the tested probes (Table 6).

4.6 Choice of Intermediate Layers

Figure 5 tests stage composition and layer sampling density. L234 slightly exceeds equal-count L123, L124, and L134 in mean accuracy on both tasks. Adding L1 or doubling the predictors through dense sampling of L2–L4 does not improve mean full model accuracy. These comparisons favor a compact layer set spanning different depths over denser sampling. This accords with Proposition 1's emphasis on complementarity to the existing combination, rather than source count alone. Fixed layer locations define the candidates, while OOF estimation determines their combination across modalities and depths. Dense sampling slightly improves independent fold model accuracy, so L234 remains a shared default rather than a unique optimum.

Table 6: Probe ablations (%) with convex aggregation.

Variant	CREMA-D		KS	
	Acc.	Macro-F1	Acc.	Macro-F1
Logistic	85.22	85.48	77.22	77.09
Linear SVM	85.22	85.59	76.55	76.16
RFF-Ridge	85.48	85.86	76.67	76.24
Ridge	**85.66**	**86.05**	**77.74**	**77.45**

5 Conclusion

LAMF combines output and intermediate layer predictions using a combination estimated from complementary OOF records, retaining the base architecture, training objective, and single model inference. Our theory explains weakened in-sample loss responses and conditions for beneficial layer predictions. Experiments across eight tasks show competitive accuracy, while fixed model analyses and ablations support improved utilization of learned modality information. Controlled comparisons further link the gains from layer predictions to the records used for combination estimation. Future work will reduce training cost and improve fold to full transfer.

AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation and debugging, and translation. Additionally, we used these tools to edit the manuscript for linguistic clarity and readability. We have reviewed all AI-assisted work, including the implementation, translations, and manuscript revisions. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

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.

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

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000.

Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

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

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721–725, 2020.

Zijie Chen, Kejing Yin, Wenfang Yao, William Kwok-wai Cheung, and Jing Qin. MASAM: Multimodal adaptive sharpness-aware minimization for heterogeneous data fusion. In *International Conference on Learning Representations*, pp. 92384–92411, 2026.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C. Mozer. Head2Toe: Utilizing intermediate representations for better transfer learning. In *Proceedings of the International Conference on Machine Learning*, volume 162, pp. 6009–6033, 2022.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20029–20038, 2023.

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.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. ReconBoost: Boosting can achieve modality reconcilement. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 19573–19597, 2024.

Chengxiang Huang, Yake Wei, Zequn Yang, and Di Hu. Adaptive unimodal regulation for balanced multimodal information acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25854–25863, 2025.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *Proceedings of the International Conference on Machine Learning*, volume 162, pp. 9226–9259, 2022.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, volume 28, pp. 427–435, 2013.

Qing-Yuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129580–129607, 2025.

Xun Jiang, Yufan Gu, Disen Hu, Yuqing Hou, Yazhou Yao, Fumin Shen, Heng Tao Shen, and Xing Xu. Multimodal learning on low-quality data with conformal predictive self-calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23041–23050, 2026.

Amr Khalifa, Michael C. Mozer, Hanie Sedghi, Behnam Neyshabur, and Ibrahim Alabdulmohsin. Layer-stack temperature scaling. *arXiv preprint arXiv:2211.10193*, 2022.

Dieter Kraft. A software package for sequential quadratic programming. Technical Report DFVLR-FB 88-28, DFVLR, 1988.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie (Yufan) 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.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. M-SENA: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 204–213, 2022.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *Proceedings of the International Conference on MultiMedia Modeling*, pp. 15–27, 2016.

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

Yang Qin, Yanglin Feng, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Deep information-balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):9384–9396, 2026.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, pp. 12695–12705, 2020.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops*, pp. 1–6, 2015.

Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22879–22888, 2025.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 52559–52572, 2024.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. Multimodal classification via total correlation maximization. In *International Conference on Learning Representations*, pp. 45101–45117, 2026.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 2016.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *Proceedings of the European Conference on Computer Vision*, pp. 386–403, 2022.

A PROOFS

For a fixed $C \times C$ class permutation matrix Π and the identity matrix I, assume $\mathbb{E}_P \operatorname{osc}(z_0(X))$ and $\mathbb{E}_P \operatorname{osc}(s_j(X))$ are finite. Substituting $a_j = s_j - z_0$ in Equation (4) and subtracting gives

$$A(s_j; z_0) = \mathbb{E}_P[(p_0(X) - e_Y)^\top z_0(X)] - \mathbb{E}_P[(p_0(X) - e_Y)^\top s_j(X)],$$
$$A(s_j; z_0) - A(\Pi s_j; z_0) = \mathbb{E}_P[(e_Y - p_0(X))^\top (I - \Pi) s_j(X)].$$

A.1 PROOF OF PROPOSITION 1

Proposition 1. Fix the model, probes, and scales. Let $a_j = s_j - z_0$ and $L_j(\eta) = \mathbb{E}_P \ell(z_0(X) + \eta a_j(X), Y)$. If $L_j(0) < \infty$ and $\mathbb{E}_P \operatorname{osc}(a_j(X)) < \infty$, then

$$A(s_j; z_0) = -L_j'(0^+) = \mathbb{E}_P[(e_Y - p_0(X))^\top a_j(X)]. \tag{4}$$

If $A(s_j; z_0) > 0$, then $L_j(\eta) < L_j(0)$ for sufficiently small $\eta > 0$; if $A(s_j; z_0) \le 0$, then $L_j(\eta) \ge L_j(0)$ for all $\eta \in [0, 1]$.

Lemma 1 (Bhatia–Davis inequality; Bhatia & Davis, 2000). If $Z \in [l, u]$ almost surely and $\mu = \mathbb{E}Z$, then $\operatorname{Var}(Z) \le (u - \mu)(\mu - l) \le (u - l)^2/4$.

Proof of Proposition 1. Since $\nabla_z \ell(z, y) = \operatorname{softmax}(z) - e_y$, the difference quotient along $a_j(x)$ is bounded in absolute value by $\operatorname{osc}(a_j(x))$. Dominated convergence therefore gives

$$\begin{aligned} L_j'(0^+) &= \lim_{\eta \downarrow 0} \mathbb{E}_P \left[\frac{\ell(z_0(X) + \eta a_j(X), Y) - \ell(z_0(X), Y)}{\eta} \right] \\ &= \mathbb{E}_P \left[\lim_{\eta \downarrow 0} \frac{\ell(z_0(X) + \eta a_j(X), Y) - \ell(z_0(X), Y)}{\eta} \right] \\ &= \mathbb{E}_P \left[\sum_c p_{0,c}(X) \big([a_j(X)]_c - [a_j(X)]_Y \big) \right] \\ &= \mathbb{E}_P[(p_0(X) - e_Y)^\top a_j(X)] \\ &= -A(s_j; z_0). \end{aligned} \tag{12}$$

A negative derivative gives a strict local decrease. If $A(s_j; z_0) \le 0$, convexity gives $L_j(\eta) \ge L_j(0) + \eta L_j'(0^+) \ge L_j(0)$. If z_0 attains $L_0^\star$, feasibility gives $G \ge L_j(0) - L_j(\eta)$.

If additionally $\kappa_j = \frac{1}{4}\mathbb{E}_P \operatorname{osc}(a_j(X))^2 < \infty$, set $p_t(x) = \operatorname{softmax}(z_0(x) + t a_j(x))$. A second differentiation, Taylor's formula with integral remainder, and Lemma 1 yield, for $0 \le \eta \le 1$,

$$\begin{aligned} L_j(0) - L_j(\eta) &= \eta A(s_j; z_0) - \int_0^\eta (\eta - t) \mathbb{E}_P \left[\operatorname{Var}_{c \sim p_t(X)} ([a_j(X)]_c) \right] dt \\ &\ge \eta A(s_j; z_0) - \kappa_j \int_0^\eta (\eta - t)\, dt \\ &= \eta A(s_j; z_0) - \tfrac{1}{2} \kappa_j \eta^2. \end{aligned} \tag{13}$$

When $A(s_j; z_0) > 0$, $\kappa_j > 0$ and $\eta_j = \min\{A(s_j; z_0)/\kappa_j, 1\}$ maximizes this lower bound. □

A.2 PROOF OF PROPOSITION 2

Proposition 2. If $\operatorname{osc}(a_j(x_i)) \le B_j < \infty$ for every training sample, with $B_j \ge 0$, then

$$|\widehat{A}_{\mathrm{tr},j}| \le \frac{B_j}{N} \sum_i (1 - p_{0,y_i}(x_i)), \qquad 0 \le \widehat{L}''_{\mathrm{tr},j}(0) \le \frac{B_j^2}{N} \sum_i (1 - p_{0,y_i}(x_i)). \tag{6}$$

Both quantities vanish as $p_{0,y_i}(x_i) \to 1$ with B_j fixed.

Proof. Set $a_i = a_j(x_i)$, $p_i = p_0(x_i)$, and $\nu_i = 1 - p_{i,y_i}$. Differentiating the empirical loss and using $|a_{i,c} - a_{i,y_i}| \le B_j$ gives

$$\begin{aligned} |\widehat{A}_{\mathrm{tr},j}| &= \Big|\frac{1}{N}\sum_i \sum_{c\neq y_i} p_{i,c}(a_{i,y_i} - a_{i,c})\Big| \\ &\le \frac{1}{N}\sum_i \sum_{c\neq y_i} p_{i,c}|a_{i,y_i} - a_{i,c}| \le \frac{B_j}{N}\sum_i \nu_i, \\ 0 \le \widehat{L}''_{\mathrm{tr},j}(0) &= \frac{1}{N}\sum_i \mathrm{Var}_{c\sim p_i}(a_{i,c}) \\ &\le \frac{1}{N}\sum_i \sum_{c\neq y_i} p_{i,c}(a_{i,c} - a_{i,y_i})^2 \le \frac{B_j^2}{N}\sum_i \nu_i. \end{aligned} \tag{14}$$

The variance bound uses deviations from a_{i,y_i}; substituting $\nu_i = 1 - p_{0,y_i}(x_i)$ gives Equation (6) and its limits.

For finite changes between $q, q' \in \Delta_J$, now set $z_i = z_q(x_i)$, $a_i = z_{q'}(x_i) - z_i$, $p_i = \mathrm{softmax}(z_i)$, and $\nu_i = 1 - p_{i,y_i}$, with $\mathrm{osc}(a_i) \le B < \infty$. Write $\widehat{L}_{\mathrm{tr}}(q) = N^{-1}\sum_i \ell(z_q(x_i), y_i)$. Jensen's inequality and $\log(1+t) \le t$ give

$$\begin{aligned} \ell(z_i + a_i, y_i) - \ell(z_i, y_i) &= \log\Big[p_{i,y_i} + \sum_{c\neq y_i} p_{i,c} e^{a_{i,c} - a_{i,y_i}}\Big], \\ -\nu_i B \le \ell(z_i + a_i, y_i) - \ell(z_i, y_i) &\le \nu_i(e^B - 1), \\ \big|\widehat{L}_{\mathrm{tr}}(q') - \widehat{L}_{\mathrm{tr}}(q)\big| &\le \frac{e^B - 1}{N}\sum_i \nu_i \le (e^B - 1)\widehat{L}_{\mathrm{tr}}(q), \end{aligned} \tag{15}$$

where the last line uses $B \le e^B - 1$ and $\nu_i \le -\log p_{i,y_i}$. □

Corollary 1 (Empirical optimality and population improvement). Fix the model, probes, and scales, with at least one layer source. Let $\mathcal{J}_0$ index the original heads, $\Delta_0 = \{q \in \Delta_J : q_j = 0 \text{ for } j \notin \mathcal{J}_0\}$, and $q_0 \in \arg\min_{q\in\Delta_0} \widehat{L}_{\mathrm{tr}}(q)$, with $z_0 = z_{q_0}$. Define $\widehat{A}_{\mathrm{tr},j}$ by Equation (5) for every source and $u_+ = \max\{u, 0\}$. Under Proposition 2's bounds for the layer sources,

$$\begin{aligned} 0 \le \widehat{L}_{\mathrm{tr}}(q_0) - \min_{q\in\Delta_J} \widehat{L}_{\mathrm{tr}}(q) &\le \max_{j\notin\mathcal{J}_0}(\widehat{A}_{\mathrm{tr},j})_+ \\ &\le \Big(\max_{j\notin\mathcal{J}_0} B_j\Big)\frac{1}{N}\sum_i (1 - p_{0,y_i}(x_i)). \end{aligned} \tag{16}$$

If some $j \notin \mathcal{J}_0$ satisfies Proposition 1's assumptions and $A(s_j; z_0) > 0$, then $L_j(\eta) < L_j(0)$ for sufficiently small $\eta > 0$.

Proof. Optimality on Δ_0 gives $\widehat{A}_{\mathrm{tr},j} \le 0$ for $j \in \mathcal{J}_0$. For every $q \in \Delta_J$, convexity yields

$$\begin{aligned} \widehat{L}_{\mathrm{tr}}(q) - \widehat{L}_{\mathrm{tr}}(q_0) &\ge \frac{1}{N}\sum_i (p_0(x_i) - e_{y_i})^\top (z_q(x_i) - z_0(x_i)) \\ &= -\sum_j q_j \widehat{A}_{\mathrm{tr},j} \\ &\ge -\sum_{j\notin\mathcal{J}_0} q_j (\widehat{A}_{\mathrm{tr},j})_+ \ge -\max_{j\notin\mathcal{J}_0}(\widehat{A}_{\mathrm{tr},j})_+. \end{aligned} \tag{17}$$

Taking the infimum over q and using Equation (6) yields Equation (16); the population statement follows from Proposition 1. Equation (13) gives a quantitative bound when $\kappa_j < \infty$. If all layer sources satisfy $\widehat{A}_{\mathrm{tr},j} \le 0$, Equation (16) gives $q_0 \in \arg\min_{q\in\Delta_J} \widehat{L}_{\mathrm{tr}}(q)$. □

A.3 Proof of Proposition 3

Proposition 3. Suppose the expected losses and discrepancies in Section 3.4 are finite, and $\widehat{L}_{\mathrm{OOF}}(\widehat{q}) \le \inf_{q\in\Delta_J} \widehat{L}_{\mathrm{OOF}}(q) + \xi$ for $\widehat{q} \in \Delta_J$ and $\xi \ge 0$. Then

$$L(\widehat{q}) - L_0^\star \le 2\epsilon + 2\Gamma + \xi - G, \qquad G = L_0^\star - L_+^\star. \tag{11}$$

Proof. Equation (10) gives $|L(q) - \widehat{L}_{\mathrm{OOF}}(q)| \le \epsilon + \Gamma$ on Δ_J. Hence

$$\begin{aligned} L(\widehat{q}) - L_0^\star &\le \widehat{L}_{\mathrm{OOF}}(\widehat{q}) - L_0^\star + \epsilon + \Gamma \\ &\le \inf_{q \in \Delta_J} \widehat{L}_{\mathrm{OOF}}(q) - L_0^\star + \epsilon + \Gamma + \xi \\ &\le L_+^\star - L_0^\star + 2\epsilon + 2\Gamma + \xi \\ &= 2\epsilon + 2\Gamma + \xi - G. \qquad \square \end{aligned} \tag{18}$$

Lemma 2 (Hoeffding's inequality; Hoeffding, 1963). For independent $Z_1, \dots, Z_n \in [0, B_\ell]$ with $B_\ell > 0$, $\Pr(|n^{-1}\sum_i (Z_i - \mathbb{E}Z_i)| > t) \le 2\exp(-2nt^2/B_\ell^2)$ for $t > 0$.

Assume independent pairs from P_{tr} with positive class probabilities; split within classes independently of features given labels. Let I_k index $\mathcal{D}_k$, $n_k = |I_k| > 0$, $\omega_k = n_{3-k}/N$, and $n_{\min} = \min(n_1, n_2)$. Let $\mathcal{B}_k$ contain all labels, the split, fold k's training features and independent training randomness.

Write $s_{j,k}(x) = s_j(x; F_k)$. Let $\mathcal{F}_k$ collect $f(x,y) = \ell(\sum_j q_j s_{j,k}(x), y)$ over the configurations, scales and weights in Section 3.3, with models and probes fitted only on $\mathcal{D}_k$. Write $\widehat{L}_k(f) = n_{3-k}^{-1}\sum_{i \in I_{3-k}} f(X_i, Y_i)$, $M_k(f) = \mathbb{E}[\widehat{L}_k(f) \mid \mathcal{B}_k]$, and $L_k(f) = \mathbb{E}_{P_{\mathrm{tr}}}[f(X,Y) \mid \mathcal{B}_k]$ for independent $(X,Y) \sim P_{\mathrm{tr}}$.

For deterministic $0 < B_\ell, \zeta, \mathcal{N}(\zeta) < \infty$, assume $0 \le f \le B_\ell$ on the support of P_{tr} and a $\mathcal{B}_k$-measurable sup-norm ζ-cover of $\mathcal{F}_k$ of size $\le \mathcal{N}(\zeta)$. Lemma 2 and a union bound give

$$\Pr\left(\sup_{f \in \mathcal{F}_k} |\widehat{L}_k(f) - M_k(f)| > 2\zeta + t \,\middle|\, \mathcal{B}_k \right) \le 2\mathcal{N}(\zeta) \exp\left(-\frac{2n_{3-k}t^2}{B_\ell^2} \right). \tag{19}$$

For $0 < \delta < 1$, take $t = B_\ell\sqrt{\log(4\mathcal{N}(\zeta)/\delta)/(2n_{\min})}$ and $\pi = (P_{\mathrm{tr}}(Y = c))_c$, $\widehat{\pi}_k = (n_k^{-1}\sum_{i \in I_k} \mathbf{1}\{Y_i = c\})_c$, $L_{k,c}(f) = \mathbb{E}[f(X,c) \mid Y = c, \mathcal{B}_k]$. Since $M_k(f) = \sum_c \widehat{\pi}_{3-k,c} L_{k,c}(f)$ and $L_k(f) = \sum_c \pi_c L_{k,c}(f)$, a union bound gives, with probability at least $1 - \delta$,

$$\begin{aligned} \epsilon &\le \sum_k \omega_k \sup_{f \in \mathcal{F}_k} |\widehat{L}_k(f) - L_k(f)| \\ &\le 2\zeta + t + \sum_k \omega_k \sup_{f \in \mathcal{F}_k} \left| \sum_c (\widehat{\pi}_{3-k,c} - \pi_c)\big(L_{k,c}(f) - B_\ell/2\big) \right| \\ &\le 2\zeta + B_\ell\sqrt{\frac{\log(4\mathcal{N}(\zeta)/\delta)}{2n_{\min}}} + \tfrac{B_\ell}{2}\sum_k \omega_k \|\widehat{\pi}_{3-k} - \pi\|_1. \quad \square \end{aligned} \tag{20}$$

A.4 Risk Comparison for Regularized Linear Stacking

Let $u = [s_1^\top, \dots, s_J^\top]^\top$ and $\widetilde{u} = D_u^{-1}(u - \mu_u)$, where μ_u contains OOF coordinate means and D_u the positive diagonal scales. For $\vartheta = (V, b_{\mathrm{agg}})$, $V \in \mathbb{R}^{C \times JC}$, and $b_{\mathrm{agg}} \in \mathbb{R}^C$,

$$\begin{aligned} z_\vartheta = V\widetilde{u} + b_{\mathrm{agg}}, \qquad \lambda > 0, \quad \xi_{\mathrm{lin}} \ge 0, \\ \widehat{L}_{\mathrm{lin}}(\widehat{\vartheta}) + \tfrac{\lambda}{2}\|\widehat{V}\|_F^2 \le \inf_\vartheta \{\widehat{L}_{\mathrm{lin}}(\vartheta) + \tfrac{\lambda}{2}\|V\|_F^2\} + \xi_{\mathrm{lin}}, \end{aligned} \tag{21}$$

where $\widehat{L}_{\mathrm{lin}}$ is mean OOF cross entropy. Centering the fitted, unpenalized intercept preserves probabilities. Set

$$\Theta = \{(V, b_{\mathrm{agg}}) : \|V\|_F \le \|\widehat{V}\|_F,\ \|b_{\mathrm{agg}}\|_\infty \le \|\widehat{b}_{\mathrm{agg}}\|_\infty,\ \mathbf{1}^\top b_{\mathrm{agg}} = 0\}. \tag{22}$$

Define $L_{\mathrm{lin}}, \overline{L}_{\mathrm{lin}}, \epsilon_{\mathrm{lin}}, \Gamma_{\mathrm{lin}}$ as in Section 3.4 over Θ, with the same map and standardizer across models. Let $\Theta_0 \subseteq \Theta$ set layer predictor coefficient blocks to zero. Write $L_{\mathrm{lin}}^\star = \inf_\Theta L_{\mathrm{lin}}$, $L_{\mathrm{out}}^\star = \inf_{\Theta_0} L_{\mathrm{lin}}$, and $G_{\mathrm{lin}} = L_{\mathrm{out}}^\star - L_{\mathrm{lin}}^\star$. For finite risks and discrepancies, the regularized objective gives, for every $\vartheta \in \Theta$,

$$\begin{aligned} \widehat{L}_{\mathrm{lin}}(\widehat{\vartheta}) &\le \widehat{L}_{\mathrm{lin}}(\vartheta) + \tfrac{\lambda}{2}(\|V\|_F^2 - \|\widehat{V}\|_F^2) + \xi_{\mathrm{lin}} \le \widehat{L}_{\mathrm{lin}}(\vartheta) + \xi_{\mathrm{lin}}, \\ L_{\mathrm{lin}}(\widehat{\vartheta}) - L_{\mathrm{out}}^\star &\le 2\epsilon_{\mathrm{lin}} + 2\Gamma_{\mathrm{lin}} + \xi_{\mathrm{lin}} - G_{\mathrm{lin}}. \end{aligned} \tag{23}$$

The risk bound follows as in Equation (18). $\square$

A.5 Conditional Risk Consistency and Accuracy Bounds

Risk consistency. If $\epsilon, \Gamma, \xi \to 0$ in probability as sample size grows, then $L(\widehat{q}) - L_+^\star \to 0$ in probability. **Proof.** Apply Equation (18); Equation (23) gives the stacker analogue. □

Lemma 3 (Pinsker's inequality; Tsybakov, 2009). For probability vectors p, r, $\|p - r\|_1 \le \sqrt{2\,\mathrm{KL}(p\|r)}$, with natural logarithms.

Let $p^*(x) = P(Y = \cdot \mid X = x)$, $p_q(x) = \mathrm{softmax}(\sum_j q_j s_j(x; F))$, and let g^*, g_q be their maximizing-class rules with fixed tie breaking. Write $R_{01}(g) = P(g(X) \neq Y)$, $\mathrm{Acc} = 1 - R_{01}$, and $H = \mathbb{E}_P[-\log p_Y^*(X)]$. For a fixed output only baseline g_0, define $a = L_+^\star - H$ and $b = R_{01}(g_0) - R_{01}(g^*) > 0$. **Accuracy gain.** If $a + 2\epsilon + 2\Gamma + \xi < b^2/2$, then $\mathrm{Acc}(g_{\widehat{q}}) > \mathrm{Acc}(g_0)$. **Proof.** The pointwise classification regret is at most $\|p^* - p_q\|_1$. Lemma 3 and Jensen's inequality give

$$\begin{aligned} R_{01}(g_q) - R_{01}(g^*) &\le \mathbb{E}_P \|p^*(X) - p_q(X)\|_1 \\ &\le \sqrt{2\mathbb{E}_P\,\mathrm{KL}(p^*(X)\|p_q(X))} = \sqrt{2\{L(q) - H\}}, \\ \mathrm{Acc}(g_{\widehat{q}}) - \mathrm{Acc}(g_0) &\ge b - \sqrt{2\{a + 2\epsilon + 2\Gamma + \xi\}}, \end{aligned} \tag{24}$$

For stacking, replace a by $L_{\mathrm{lin}}^\star - H$ and use the corresponding discrepancies. □

B Experimental Details and Cost

B.1 Datasets and Inputs

Dataset statistics are in Table 1; images and video frames use 224×224 resolution.

Audio–visual inputs. The four audio–visual tasks use separate audio and visual ResNet18 encoders trained from scratch. Audio is represented by log-magnitude STFT spectrograms, following the input pipeline of OGM-GE (Peng et al., 2022). Training frames use random resized crops and horizontal flips; evaluation frames are resized, with ImageNet channel normalization.

CREMA-D (Cao et al., 2014) contains six emotion classes. We repeat or truncate audio to 20 seconds and use a fixed single frame for both training and evaluation, with visual preprocessing adapted from OGM-GE (Peng et al., 2022).

Kinetics-Sounds (Arandjelović & Zisserman, 2017) uses a 33-category subset with paired audio and video. Each input contains 10 seconds of audio and three uniformly sampled frames.

AVSBench and VGGSound50. AVSBench is used for clip classification (Zhou et al., 2022); VGGSound50 contains 50 categories from VGGSound (Chen et al., 2020). Following IBML's three-second audio and three-segment sampling design (Qin et al., 2026), our adapter orders frames temporally and selects one per segment, randomly during training and at segment centers during evaluation.

UPMC-Food101 (Wang et al., 2015) pairs food images with text. We train a ResNet18 image encoder and a bidirectional GRU text encoder from scratch, including the word embeddings. Text uses lowercased WordPiece tokenization with a vocabulary of up to 16,000 tokens fitted on training text. Images follow the crop/flip and resize pipeline above.

MVSA-Single (Niu et al., 2016) is a three-class sentiment task. We fine-tune pretrained BERT-base-uncased and ResNet152 with ImageNet-1K V1 weights. Following the CPSC input implementation (Jiang et al., 2026), text uses BERT tokenization with at most 512 tokens; images are resized to a shorter side of 256 and center-cropped to 224.

CMU-MOSI and CMU-MOSEI (Zadeh et al., 2016; Bagher Zadeh et al., 2018) use fixed, aligned features released by MMSA (Mao et al., 2022), with sequence length 50 and pre-extracted BERT text representations. Three independent Transformer branches follow a MultiBench-style configuration (Liang et al., 2021) and are trained from scratch on these features; pooling excludes padded positions. Has0 retains neutral samples and distinguishes positive sentiment from non-positive sentiment.

B.2 Baselines and Reporting

Unimodal denotes independently trained models; other modality columns report branch predictions. Linear concatenation heads are decomposed into modality-specific coefficient blocks with equally divided bias, consistent with TCMax's bimodal implementation (Yu et al., 2026). Convex aggregation and linear stacking are reported as separate implementations, with branch results taken from each implementation's selected model state. Their main table difference is therefore not an isolated aggregator effect; fixed model comparisons are reported separately. Unless stated otherwise, results are three-seed means $\pm$ sample standard deviations, with differences paired within seeds. Seeds are 1001–1003 except for MOSI (1004–1006) and Figure 1(b) (9023–9026); Table 4 uses seed 1001. Accuracy changes are in percentage points (pp); cross entropy (CE) is in nats.

Table 7 summarizes parameter estimation. OOF labels select configurations and scales as well as the combination; they are not an independent evaluation set for these choices.

Table 7: Records used for parameter fitting and selection in LAMF.

Stage	Estimated quantities	Records and objective
Base training	Encoders and original heads	Fit on the model's training subset using its original objective.
Probe fitting	Feature statistics, coefficients and intercepts	Fit on the same subset as the corresponding model.
Probe configuration	Pooling and ridge penalty	Select from pooled OOF records using the task criterion.
Score scales	Positive probe scales	Minimize each probe's OOF CE; original heads keep unit scale.
Convex combination	Global simplex coefficients	Minimize OOF CE with configurations and scales fixed.
Linear stacking	Score statistics, penalty and linear map	Select regularization within OOF records, then fit the final map.
Deployment	Full model probes	Refit on full training representations; retain configurations, scales and the combination.

B.3 Implementation Settings

Base model training. LAMF uses independent branch classifiers and an MLP fusion head on concatenated representations. The objective sums fusion CE, with coefficient one, and weighted branch CEs (Table 8). Fold and full models share the architecture and objective; the frozen models in Table 4 retain their original methods. Learning rates and training durations follow the task and backbone: LAMF constructs layer predictions and estimates their combination from OOF records, rather than introducing a training optimizer that requires a shared learning rate or epoch count.

Table 8: Base model training settings. The final column gives each branch CE coefficient; MOSI lists convex/stacking settings.

Dataset	Optimizer	Initial LR	Epochs	Batch	Branch CE
CREMA-D	SGD	0.01	200	64	1
Kinetics-Sounds	SGD	0.001	150	64	4
AVSBench	SGD	0.001	100	64	2
VGGSound50	SGD	0.001	100	64	4
UPMC-Food101	SGD	0.001	100	64	1
MVSA-Single	Adam	0.00003	50	32	4
CMU-MOSI	SGD	0.001	30	64	$1, \frac{1}{3}$
CMU-MOSEI	SGD	0.001	100	64	1

SGD uses momentum 0.9 and weight decay 10^{-4}; Adam uses no weight decay. CREMA-D and KS multiply the learning rate by 0.1 after epochs 160/180 and 100/135, respectively; AVSBench and VGGSound50 do so after epoch 70. Food-101 starts at 0.001, warms up to 0.01, resets to 0.001 at epoch 37, and uses 0.0001 from epoch 72. MVSA accumulates 24 minibatches and halves the rate after two epochs without improvement in validation fusion accuracy. Sentiment tasks use a constant rate; MOSI fold training matches full model update counts. KS freezes the audio encoder after epoch 75 while continuing to train its normalization layers, heads, visual encoder and fusion module.

Prediction sources for image–text tasks. Both tasks combine three original heads and six probes. Food-101 text probes use masked temporal means of embeddings, forward GRU states, and direction-averaged bidirectional states; MVSA uses BERT CLS representations at layers 4, 8, and 12. Image probes use globally averaged ResNet stages 2, 3, and 4. Stage labels denote these architecture-specific locations.

Prediction sources for sentiment tasks. In Table 3, MOSI and MOSEI combine the fusion head with three ridge predictors fitted to fixed input features, one per modality; these predictors replace the branch heads in the prediction set. The Transformer layer additions in Appendix E.3 are a separate diagnostic.

Probe fitting. Folds are class-stratified, grouping sentiment clips by source video. Standardization and probe fitting use the model's training subset. Bimodal ridge penalties range over $\{10^{-6}, 10^{-5}, \ldots, 10^{2}\}$, selected by OOF accuracy with ties favoring stronger regularization. Probe scales minimize OOF CE over $\log \gamma_j \in [-4, 6]$; original heads retain unit scale.

Probe alternatives. Table 6 substitutes regularized multinomial logistic regression, one-versus-rest squared-hinge Linear SVM, or ridge on fixed 512-dimensional random Fourier features (Rahimi & Recht, 2007) with bandwidth from standardized training features. Layers and convex aggregation remain fixed.

Combination estimation. SLSQP (Kraft, 1988) solves the simplex objective in Equation (8). Linear stacking uses L-BFGS (Liu & Nocedal, 1989) and L_2 regularization selected by CE in stratified five-fold cross-validation of fixed OOF records. Each split fits its own score standardizer and stacker, not encoders or probes.

B.4 Computational Cost

Table 9 reports model training costs. Cached fitting includes ridge fitting, configuration and scale estimation, full model probe refitting, and convex optimization for Output only and +Both. Timings are medians of five runs on one cache per task and exclude feature extraction, data loading, and stacking.

Table 9: Training cost (GPU h), total/full model ratio, and cached fitting time (s).

Dataset	Full model	Folds	Total	Ratio	CPU (s)
CREMA-D	1.447 ± 0.014	1.452 ± 0.008	2.900 ± 0.017	2.003 ± 0.011	18.76
Kinetics-Sounds	5.345 ± 0.053	5.360 ± 0.278	10.706 ± 0.279	2.003 ± 0.054	55.32

Deployment adds $C \sum_{j \notin \mathcal{J}_0} (d_j + 1)$ probe coefficients, standardizers, scales, and either J convex coefficients or $JC^2 + C$ stacking coefficients. It requires no additional encoder pass. At $J = 9$, the gate in Appendix D.2 has 252 parameters, compared with eight free coefficients for convex aggregation.

C Supplementary Comparisons

C.1 Additional Metrics and Variability

Tables 10–11 supplement the main results with LAMF Macro-F1 and accuracy standard deviations for all methods.

Table 10: LAMF accuracy and Macro-F1 (%).

	Convex		Linear stacking	
Dataset	Accuracy	Macro-F1	Accuracy	Macro-F1
CREMA-D	85.66 ± 0.54	86.05 ± 0.52	86.38 ± 0.47	86.70 ± 0.55
Kinetics-Sounds	77.74 ± 0.73	77.45 ± 1.01	78.81 ± 0.28	78.69 ± 0.37
AVSBench	92.39 ± 0.21	90.42 ± 0.06	92.21 ± 0.56	90.88 ± 0.63
VGGSound50	60.61 ± 0.28	59.57 ± 0.24	62.18 ± 0.71	62.08 ± 0.71
UPMC-Food101	92.08 ± 0.07	92.02 ± 0.06	91.24 ± 0.06	91.21 ± 0.06
MVSA-Single	79.58 ± 0.51	75.19 ± 1.09	81.70 ± 0.58	78.11 ± 0.35
CMU-MOSI	80.56 ± 0.30	79.71 ± 0.29	80.23 ± 0.43	79.45 ± 0.79
CMU-MOSEI	75.90 ± 0.15	75.90 ± 0.15	76.02 ± 0.23	76.01 ± 0.23

Table 11: Baseline fusion accuracy (%) with sample standard deviations.

Method	CREMA-D	Kinetics-Sounds	AVSBench	VGGSound50
Concat	68.55 ± 0.36	61.30 ± 0.65	83.51 ± 1.24	50.47 ± 0.19
OGM-GE	76.66 ± 0.51	69.30 ± 0.48	88.15 ± 0.28	54.81 ± 0.10
MMPareto	73.61 ± 1.01	73.10 ± 0.43	90.81 ± 0.54	58.54 ± 0.24
InfoReg	77.64 ± 0.31	73.85 ± 0.74	91.71 ± 0.28	59.90 ± 0.28
DGL	81.41 ± 0.64	77.51 ± 0.62	94.14 ± 0.08	61.24 ± 0.52
AUG	84.41 ± 0.40	77.67 ± 0.78	88.02 ± 0.43	58.06 ± 0.08
TCMax	83.51 ± 0.28	76.12 ± 0.21	87.16 ± 0.94	60.40 ± 0.16
IBML	80.06 ± 1.20	77.44 ± 0.48	92.34 ± 0.39	60.54 ± 0.57
CPSC	83.51 ± 0.76	74.69 ± 0.23	87.30 ± 0.47	56.68 ± 0.54
Method	UPMC-Food101	MVSA-Single	CMU-MOSI	CMU-MOSEI
Concat	87.73 ± 0.03	71.55 ± 2.94	78.57 ± 0.44	75.18 ± 0.14
OGM-GE	87.87 ± 0.10	78.74 ± 0.78	78.67 ± 0.72	75.08 ± 0.13
MMPareto	90.26 ± 0.07	77.59 ± 0.62	79.83 ± 0.51	75.44 ± 0.36
InfoReg	90.29 ± 0.12	80.22 ± 0.44	79.64 ± 0.94	74.85 ± 0.22
DGL	90.07 ± 0.14	78.16 ± 2.13	78.86 ± 0.39	75.05 ± 0.13
AUG	90.45 ± 0.04	80.59 ± 0.64	78.23 ± 1.18	75.01 ± 0.35
TCMax	90.19 ± 0.05	80.33 ± 0.18	78.81 ± 0.08	75.55 ± 0.16
IBML	90.67 ± 0.11	78.36 ± 0.91	79.25 ± 1.32	74.89 ± 0.32
CPSC	89.41 ± 0.13	79.13 ± 0.62	78.47 ± 0.51	75.20 ± 0.26

Paired main table comparisons. Table 12 fixes the full models and test samples of three linear stacking rows in Table 3. Original uses the native fusion head; Output only fits stacking to the original heads with the same OOF records, regularization, and standardization procedures. Output gain is Output only minus Original; set gain is LAMF minus Output only, both paired within seeds. Food-101 and MVSA add six intermediate layer predictors; MOSEI's set gain measures replacement of branch predictions with input-feature predictions, retaining the fusion head.

Table 12: Paired linear stacking comparisons: accuracy (%) and gains (pp), mean ± sample standard deviation over three seeds.

Dataset	Original	Output only	LAMF	Output gain	Set gain
UPMC-Food101	89.26 ± 0.02	90.02 ± 0.06	91.24 ± 0.06	0.76 ± 0.05	1.22 ± 0.10
MVSA-Single	68.21 ± 0.19	78.10 ± 1.24	81.70 ± 0.58	9.89 ± 1.06	3.60 ± 1.28
CMU-MOSEI	73.30 ± 1.29	72.43 ± 0.35	76.02 ± 0.23	-0.87 ± 1.63	3.59 ± 0.41

C.2 Comparison with LATES

The LATES adaptation (Khalifa et al., 2022) combines three heads and six CE-trained probes at LAMF's pooled layer positions. Projected SGD fits nonnegative coefficients on 670 excluded CREMA-D training examples or 2,318 independent KS validation examples, optionally refined by L-BFGS-B. LAMF uses OOF convex estimation; both share the KS full model. Probe loss, constraints, and records vary together here, whereas Appendix D.1 varies records within each combination family.

Table 13: Paired LAMF minus LATES differences. Positive accuracy and negative CE differences favor LAMF.

Dataset	LATES adaptation	Δ Accuracy	Δ Cross entropy
CREMA-D	Projected SGD	2.33 ± 1.40	-0.0480 ± 0.0172
	Solver refinement	2.33 ± 1.02	-0.0345 ± 0.0192
Kinetics-Sounds	Projected SGD	0.45 ± 0.22	-0.0274 ± 0.0022
	Solver refinement	0.41 ± 0.42	-0.0261 ± 0.0014

D Ablations and Design Choices

D.1 Estimation Records and Selection Sensitivity

We fix layer positions, OOF-selected ridge configurations and scales, full model predictions, and test samples, then refit each combination, including stacking regularization. In-sample conditions use the records on which their source models were fitted, whereas OOF uses the opposite halves. Calibration uses 670 excluded CREMA-D training examples or 2,318 independent KS validation examples that were not used to fit the source model. Output only combines three heads, and +Both adds six probes.

Table 14: Paired test accuracy differences (pp), OOF minus the alternative estimation records.

Alternative	Combination	CREMA-D		Kinetics-Sounds	
		Output only	+Both	Output only	+Both
Full in-sample	Convex	12.99 ± 1.70	14.24 ± 1.29	7.33 ± 0.57	4.32 ± 0.46
Fold in-sample	Convex	13.22 ± 1.48	14.47 ± 1.40	3.17 ± 0.89	4.34 ± 0.68
Calibration	Convex	0.48 ± 0.32	0.69 ± 0.58	0.38 ± 0.23	0.43 ± 0.52
Full in-sample	Linear stacking	9.09 ± 1.85	8.38 ± 0.79	3.74 ± 0.46	2.88 ± 0.53
Fold in-sample	Linear stacking	9.86 ± 1.56	7.80 ± 0.82	4.43 ± 1.48	2.96 ± 0.64
Calibration	Linear stacking	1.48 ± 0.59	2.42 ± 1.50	3.37 ± 0.86	1.98 ± 0.15

When calibration also selects probe penalties and scales before full model refitting, OOF improves accuracy by 0.43 ± 0.38 pp on CREMA-D and 0.35 ± 0.45 pp on KS. We also match OOF record

counts to 670 and 2,318, respectively, using ten stratified subsets per seed. After averaging subsets within seeds, +Both gains 0.74 pp on CREMA-D and 0.45 pp on KS over Output only.

Selection reuse. With fold models fixed, we divide opposite-half records into S and E by speaker on CREMA-D or source video on KS. Condition A selects configurations and scales on S, whereas B uses $S \cup E$. Both fit combinations on S and evaluate on E before exchanging subsets. Only B uses evaluation labels for selection; neither uses them for combination fitting.

Record count. We subsample fixed OOF records by class and refit only the nine source convex combination; configurations, scales, and test scores remain fixed. Five repetitions per fraction are averaged within each seed before subtracting its 100% result.

Table 15: OOF selection sensitivity: accuracy change B minus A; CE reduction A minus B.

Dataset	Acc. (pp)	CE (nats)
CREMA-D	-0.035 ± 0.121	-0.00010 ± 0.00044
Kinetics-Sounds	-0.002 ± 0.010	0.00003 ± 0.00012

Table 16: Paired test accuracy changes (pp) relative to using 100% of OOF records.

Records	CREMA-D	KS
25%	-0.09 ± 0.40	-0.04 ± 0.25
50%	-0.09 ± 0.22	-0.04 ± 0.07
75%	-0.09 ± 0.22	-0.01 ± 0.07
100%	0.00	0.00

Numerical checks. All 108 fits comparing estimation records converged, with a maximum convex Frank–Wolfe gap (Jaggi, 2013) of 1.40×10^{-7}. Tighter solver tolerances changed the loss by at most 2.87×10^{-9} without changing test classifications. Appendix E.2 separately crosses prediction sets and estimation records on fixed fold models with an independent evaluation subset.

D.2 Score Scales and Combination Alternatives

Table 17 uses fixed full models, paired within seeds. All three scale settings use convex aggregation: Unit fixes all scales at one; Probes only fits each probe scale separately and keeps original heads at unit scale; All sources fits each source scale separately. Nonnegative jointly fits $\beta_j \geq 0$ on unscaled scores without a unit-sum constraint. Configurations and deployment predictions are fixed; OOF/calibration records follow Appendix D.1.

For fixed positive $\widehat{\gamma}_j$ (one for original heads), write $s_j = \widehat{\gamma}_j r_j$. The inverse maps

$$w_j = \frac{q_j \widehat{\gamma}_j}{\sum_k q_k \widehat{\gamma}_k}, \qquad q_j = \frac{w_j / \widehat{\gamma}_j}{\sum_k w_k / \widehat{\gamma}_k}$$

give a bijection on Δ_J, with $\sum_j q_j s_j = (\sum_j q_j \widehat{\gamma}_j) \sum_j w_j r_j$. The attainable classification decisions are unchanged under fixed tie breaking, but probabilities and cross entropy need not be.

All 24 nonnegative fits have projected KKT residuals below 3.0×10^{-7}. Their paired layer CE reductions are positive in all 12 seed/record conditions; accuracy decreases in one CREMA-D OOF seed.

For the OOF conditions in Table 17, Table 18 reports $\Delta\text{Metric} = \text{Metric}(\text{Output only}) - \text{Metric}(\text{+Both})$, averaged over three paired seeds with sample standard deviations. ECE15 uses maximum-class confidence in 15 equal-width bins on $[0, 1]$, including 1 in the final bin. Brier averages the sum of squared class-probability errors per sample, without dividing by the number of classes. Scales in All sources are fitted separately, not jointly; their ECE changes are negative.

Input dependent combinations. With predictions, configurations, and scales fixed, an affine softmax gate maps each source's normalized entropy, maximum probability, and top-two probability gap to weights. We select L_2 regularization from $\{10^{-4}, 10^{-2}, 1\}$ using group-disjoint three-fold OOF validation, with standardizers fitted within each split. The Mean gate variant deploys the average OOF weights.

Table 17: Matched scale diagnostics on fixed full models. Accuracy change is +Both minus Output only (pp); CE reduction is Output only minus +Both (nats).

Dataset	Scale/combination	Records	Acc. change	CE reduction
CREMA-D	Unit	OOF	0.36 ± 0.31	0.0279 ± 0.0093
	Probes only	OOF	1.25 ± 0.81	0.0478 ± 0.0116
	All sources	OOF	-0.13 ± 0.36	-0.0027 ± 0.0005
	Nonnegative	OOF	1.16 ± 1.25	0.0331 ± 0.0061
	Probes only	Calibration	1.04 ± 0.53	0.0447 ± 0.0136
	Nonnegative	Calibration	0.90 ± 0.74	0.0338 ± 0.0028
KS	Unit	OOF	0.02 ± 0.25	0.0863 ± 0.0033
	Probes only	OOF	1.17 ± 0.38	0.1083 ± 0.0061
	All sources	OOF	0.35 ± 0.74	-0.0030 ± 0.0011
	Nonnegative	OOF	0.37 ± 0.22	0.0241 ± 0.0051
	Probes only	Calibration	1.12 ± 0.31	0.1055 ± 0.0076
	Nonnegative	Calibration	0.45 ± 0.18	0.0238 ± 0.0043

Table 18: Paired calibration changes after adding audio and visual layer predictions. Values are in the original metric units, not percentages; positive values indicate improvement.

Dataset	Scale fitting scope	ΔECE15 $\uparrow$	ΔBrier $\uparrow$
CREMA-D	Unit	$+0.0180 \pm 0.0046$	$+0.0097 \pm 0.0029$
CREMA-D	Probes only	$+0.0126 \pm 0.0073$	$+0.0202 \pm 0.0047$
CREMA-D	All sources	-0.0044 ± 0.0051	$+0.0004 \pm 0.0004$
KS	Unit	$+0.0594 \pm 0.0031$	$+0.0130 \pm 0.0010$
KS	Probes only	$+0.0526 \pm 0.0135$	$+0.0189 \pm 0.0021$
KS	All sources	-0.0069 ± 0.0075	$+0.0008 \pm 0.0001$

Table 19: Alternatives to global convex combination: accuracy is alternative minus global; CE reduction is global minus alternative.

Prediction set	Combination	CREMA-D		Kinetics-Sounds	
		Acc. change (pp)	CE reduction	Acc. change (pp)	CE reduction
Output only	Uniform	-7.66 ± 1.07	-0.5371 ± 0.0273	-1.43 ± 0.41	-0.0994 ± 0.0043
	Gate	-0.09 ± 0.21	0.0057 ± 0.0010	-0.10 ± 0.15	0.0019 ± 0.0035
	Mean gate	0.18 ± 0.21	0.0014 ± 0.0014	-0.04 ± 0.07	-0.0006 ± 0.0010
+Both	Uniform	-5.46 ± 0.13	-0.1167 ± 0.0038	-2.40 ± 0.85	-0.0956 ± 0.0020
	Gate	-0.13 ± 0.23	0.0012 ± 0.0003	0.06 ± 0.32	0.0038 ± 0.0024
	Mean gate	-0.09 ± 0.08	-0.0001 ± 0.0002	-0.06 ± 0.21	-0.0013 ± 0.0006

D.3 Layer Selection

All prediction sets retain the three original heads. Table 20 compares each layer set with L234 on the full models. For Table 21, we fit probes on each fold's training half and divide the opposite half between estimation and evaluation. Configurations, scales, and combinations are selected on the estimation subset. We then exchange subsets so each sample is evaluated once per seed.

The default reads each stage's last block; first-block controls retain dimensions, and dense controls read both blocks. The L4 control retains nine ridge penalties per modality; concatenation fits one probe to all six L234 representations. J counts all prediction sources.

Table 20: Full model convex accuracy changes from L234 (pp).

Layer set	J	CREMA-D	KS
L2	5	-0.22 ± 0.64	0.00 ± 0.43
L3	5	-0.27 ± 0.48	-0.10 ± 0.25
L4	5	-1.12 ± 0.79	-0.80 ± 0.38
L23	7	-0.22 ± 0.47	-0.02 ± 0.09
L24	7	-0.04 ± 0.21	-0.14 ± 0.35
L34	7	-0.31 ± 0.79	-0.08 ± 0.15

Table 21: Independent fold model accuracy (%), convex aggregation.

Prediction set	J	CREMA-D	KS
Output only	3	79.09 ± 0.48	69.40 ± 0.34
L4	5	79.77 ± 0.61	70.37 ± 0.37
L4 (nine/modality)	21	79.93 ± 0.50	70.44 ± 0.39
L234 (concat)	4	79.98 ± 0.62	70.40 ± 0.40
L234 (first blocks)	9	80.46 ± 0.72	70.59 ± 0.27
L1234	11	80.46 ± 0.67	70.78 ± 0.30
L234 (all blocks)	15	80.53 ± 0.66	70.82 ± 0.34
L1234 (all blocks)	19	80.49 ± 0.67	70.82 ± 0.34
L234 (default)	9	80.40 ± 0.69	70.78 ± 0.30

E Supplementary Mechanism Analysis

E.1 Correction Opportunities and Realization

Figure 1(a) reports standalone probe accuracy on certified records from Appendix E.3's independent evaluations; $1/C$ denotes uniform guessing.

Classification certificates. Let $\mathcal{J}_0$ index the original heads and $\Delta_0 = \{q \in \Delta_J : q_j = 0 \text{ for } j \notin \mathcal{J}_0\}$. Fix (x, y). For $c \neq y$, define the class margins and common-incorrect-class margin by

$$m_{cj} = [s_j(x)]_y - [s_j(x)]_c, \qquad d = \max_{c \neq y} \min_{j \in \mathcal{J}_0} (-m_{cj}). \tag{25}$$

If c_* attains this maximum, every $q \in \Delta_0$ satisfies $[z_q(x)]_{c_*} - [z_q(x)]_y = -\sum_j q_j m_{c_* j} \geq d$. Thus $d > 0$ certifies failure of every convex combination of the original heads, as required for Figure 1(a).

The largest attainable true class margin and its dual are

$$\begin{aligned} \rho &= \max_{q \in \Delta_0} \min_{c \neq y} \sum_{j \in \mathcal{J}_0} q_j m_{cj} \\ &= \min_{v \in \Delta_{C-1}} \max_{j \in \mathcal{J}_0} \sum_{c \neq y} v_c m_{cj}, \qquad v = (v_c)_{c \neq y}. \end{aligned} \tag{26}$$

The second equality follows from Sion's minimax theorem (Sion, 1958) on the two compact simplices. Equivalently, the primal maximizes t subject to $q \in \Delta_0$ and $\sum_j q_j m_{cj} \geq t$ for every $c \neq y$. For any feasible q, v,

$$\min_{c \neq y} \sum_{j \in \mathcal{J}_0} q_j m_{cj} \leq \rho \leq \max_{j \in \mathcal{J}_0} \sum_{c \neq y} v_c m_{cj}. \tag{27}$$

A positive primal margin certifies a uniquely correct prediction; a negative dual bound certifies impossibility under any tie rule. Taking v concentrated on c_* gives $\rho \leq -d$, recovering the sufficient criterion above. The case $\rho = 0$ is not certified. □

Table 22 reports the frequencies of the common-incorrect-class subset $\mathcal{W}$ and the LP-certified subset $\mathcal{U}$, where $\mathcal{W} \subseteq \mathcal{U}$. We normalize margins by the larger of one and the maximum absolute class-centered head score. Membership requires $d > 10^{-6}$ for $\mathcal{W}$ or a dual bound below -10^{-6} for $\mathcal{U}$; boundary cases remain uncertified.

Table 23 uses the fixed models from Appendix D.2. We pool counts across three seeds; total contributions use the difference between reported accuracies, while subset contributions normalize net corrections by full test size. Rest contains all records outside $\mathcal{U}$, including uncertified boundary cases. The certificate applies only to convex combinations, so Output only stacking can correctly classify some certified records.

At least one probe is correct on 114 CREMA-D and 183 KS records in $\mathcal{U}$; convex +Both also corrects two KS records with no individually correct probe. Within $\mathcal{W}$, net corrections for convex aggregation and stacking are 3 and 9 on CREMA-D, and 7 and 38 on KS, respectively.

Table 22: Certified-subset frequencies, pooling three seeds. Counts are retained/all model–sample records.

Evaluation	Subset	CREMA-D		Kinetics-Sounds	
		Count	Rate (%)	Count	Rate (%)
Figure 1(a)	$\mathcal{W}$	1,717/20,094	8.54	9,570/62,559	15.30
Full model test	$\mathcal{W}$	150/2,232	6.72	531/4,899	10.84
Full model test	$\mathcal{U}$	184/2,232	8.24	635/4,899	12.96

Table 23: Corrections (Fixed) and regressions (Broken) from Output only to +Both. Contribution is the change in total test accuracy (pp, mean $\pm$ std).

Combination	Subset	CREMA-D			Kinetics-Sounds		
		Fixed	Broken	Contribution	Fixed	Broken	Contribution
Convex	$\mathcal{U}$	4	0	0.18 ± 0.21	11	0	0.22 ± 0.07
	Rest	46	22	1.08 ± 1.01	132	86	0.94 ± 0.43
	All	50	22	1.25 ± 0.81	143	86	1.17 ± 0.38
Linear stacking	$\mathcal{U}$	18	4	0.63 ± 0.16	61	12	1.00 ± 0.57
	Rest	54	35	0.85 ± 0.79	169	131	0.78 ± 0.49
	All	72	39	1.48 ± 0.94	230	143	1.78 ± 0.16

E.2 Training Responses and Combination Estimation

Write $z_a = s_1, z_v = s_2, z_f = s_3$ for audio, visual, and fusion head outputs. Figure 1(b) evaluates 101 equally spaced $\alpha \in [0, 1]$ in $\alpha z_a + (1-\alpha) z_v$. For mean accuracy across four CREMA-D seeds, the 0.1-pp near-optimal training/test intervals are $[0.27, 1.00]/[0.29, 0.30]$, of widths 0.73/0.01. Test records are diagnostic only. In a separate three-seed CREMA-D full model check, audio/visual accuracy is 99.97%/89.00% on training data and 68.33%/73.34% on test data; audio's mean training true class probability is 0.9997.

Table 24 evaluates full training sets at the final saved full model states. It reports the modality head with the highest mean training accuracy; fusion confidence is the original fusion head's mean true class probability, distinct from Figure 4(a,b)'s branch-average baseline.

Table 24: Training set accuracy (%) and fusion confidence at training completion, averaged over three seeds.

Dataset	Branch	Branch acc.	Fusion acc.	Fusion confidence
CREMA-D	Audio	99.97	99.97	0.9997
Kinetics-Sounds	Audio	98.94	99.98	0.9986
AVSBench	Audio	100.00	100.00	0.9996
VGGSound50	Audio	99.23	99.95	0.9974
UPMC-Food101	Text	97.86	99.78	0.9959
MVSA-Single	Image	99.87	99.87	0.9982
CMU-MOSI	Text	96.16	98.18	0.9584
CMU-MOSEI	Text	99.65	99.86	0.9973

Figure 4(a,b) uses $z_0 = (z_a + z_v)/2$ and direction $z_a - z_0$ each epoch, one seed per task, without smoothing. Absolute sample derivatives and softmax variances are averaged within folds, then weighted by sample count. Panels 4(c,d) report three-seed OOF coefficients; a2–a4/v2–v4 denote $s_4, \ldots, s_6/s_7, \ldots, s_9$. Table 25 instead uses six layer directions $a_j = s_j - z_0$ on the same fold

models and in-sample and OOF records, averaging three layers per modality within seeds before reporting three-seed means and sample standard deviations.

Table 25: Layer prediction responses on fixed fold models. Initial decrease estimates A as the negative mean slope; magnitudes are averaged samplewise.

Dataset	Modality	Records	Initial decrease	Mean |slope|	Mean curvature
CREMA-D	Audio	In-sample	-0.0055 ± 0.0011	0.0056 ± 0.0010	0.0293 ± 0.0044
	Audio	OOF	0.4099 ± 0.0147	0.6336 ± 0.0104	0.3701 ± 0.0045
	Visual	In-sample	-0.0072 ± 0.0017	0.0073 ± 0.0016	0.0434 ± 0.0084
	Visual	OOF	0.8165 ± 0.0126	0.9604 ± 0.0071	0.6676 ± 0.0157
KS	Audio	In-sample	-0.0070 ± 0.0003	0.0081 ± 0.0004	0.0360 ± 0.0007
	Audio	OOF	0.4977 ± 0.0093	0.8601 ± 0.0084	0.9826 ± 0.0026
	Visual	In-sample	-0.0059 ± 0.0005	0.0093 ± 0.0002	0.0471 ± 0.0011
	Visual	OOF	0.6263 ± 0.0118	0.9536 ± 0.0096	1.1096 ± 0.0049

The response separation persists after normalization by each direction's maximum oscillation over in-sample and OOF records. These checks condition on selected configurations. Proposition 2 bounds responses under high confidence fitting and bounded score corrections, not the correctness of the minimizing combination; conditional OOF estimation is treated in Appendix A.3.

Estimation records and realized gains. We fix both fold models at training completion and use their fitting subsets T and disjoint opposite-half subsets S and E. Within each opposite half, S and E are separated by speaker on CREMA-D and source video on KS. Probe configurations and scales are selected only on S and shared by all conditions. We fit convex combinations on either T (in-sample) or S (OOF), using three original heads or those heads plus six layer predictors. All four conditions are evaluated on the same E, which is excluded from the current selection and fitting.

Table 26: Independent evaluation of prediction sets and estimation records. Accuracy is in percent; the paired layer gain is +Both minus Output only (pp).

Dataset	Estimation records	Output only	+Both	Layer gain
CREMA-D	In-sample	64.266 ± 1.139	64.497 ± 1.101	$+0.231 \pm 0.201$
CREMA-D	OOF	78.257 ± 0.383	79.954 ± 0.775	$+1.697 \pm 0.694$
KS	In-sample	67.037 ± 0.585	67.037 ± 0.585	0.000 ± 0.000
KS	OOF	69.256 ± 0.291	70.759 ± 0.284	$+1.503 \pm 0.203$

Each seed pools its two folds' nonoverlapping evaluation records: 3,319 for CREMA-D and 10,427 for KS. These are retrospective fold model diagnostics, not the full model results in the main tables.

The OOF layer gains in Table 26 are positive in all three seeds per task. Descriptive paired 95% t intervals are $[-0.026, 3.421]$ pp for CREMA-D and $[0.999, 2.006]$ pp for KS; the CREMA-D interval includes zero. In this batch, in-sample nine source coefficients concentrate on the fusion head, with mean weights 0.885 and 0.822, respectively; the six layer predictors together receive 0.009 and approximately zero. This check does not establish which modality controls the fusion head, and equal accuracies alone do not establish identical combinations.

Confidence and class-coordinate controls. Table 27 uses the OOF Output only combination as its common reference. Temperature scaling fits one positive scalar on S with the combination fixed. The permutation control applies each of five fixed cyclic class shifts to all six layer scores, preserving their per-sample norms and entropies but changing class alignment. It then refits the combination on S. The unit-scale condition instead sets probe scales to one and refits the combination. We average permutation results within each seed before reporting means and sample standard deviations across three seeds.

Table 27: Changes relative to the OOF Output only reference. Accuracy change is control minus reference (pp); CE reduction is reference minus control (nats).

Dataset	Control	Acc. change	CE reduction
CREMA-D	Temperature only	0.000 ± 0.000	0.0323 ± 0.0020
CREMA-D	Permuted layer scores	-0.185 ± 0.396	0.0276 ± 0.0030
CREMA-D	Unit-scale layer scores	$+0.442 \pm 0.585$	0.0442 ± 0.0046
CREMA-D	Layer scores (fitted scales)	$+1.697 \pm 0.694$	0.0730 ± 0.0081
KS	Temperature only	0.000 ± 0.000	0.1086 ± 0.0024
KS	Permuted layer scores	-0.100 ± 0.106	0.1014 ± 0.0014
KS	Unit-scale layer scores	$+0.192 \pm 0.116$	0.1153 ± 0.0016
KS	Layer scores (fitted scales)	$+1.503 \pm 0.203$	0.1593 ± 0.0028

Positive temperature scaling leaves classifications unchanged. The tested permutations reduce CE without reproducing the mean accuracy gain of aligned layer predictions; removing fitted scales also reduces that gain. From the in-sample nine source combination, all 72 actual layer and 360 permuted directions have negative initial CE derivatives on E. The derivative includes a shared confidence-scaling term (Appendix A), so its sign alone does not identify independent class information.

Gradient intervention. Table 28 scales encoder gradients before global clipping for one CREMA-D seed; head gradients remain unscaled. Conditions share full/fold initializations, splits, sample order, architecture, objective, schedule, and the ridge/scale/OOF convex procedure. Fitting gap change is relative to $(1, 1)$ for the audio-minus-visual true class probability integrated over normalized training time on a fixed training subset. Layer gains compare +Both with Output only within each condition.

Table 28: Single-seed CREMA-D gradient intervention: fitting gap changes and layer gains.

Audio/visual multipliers	Fitting gap change (pp)	Acc. gain (pp)	CE reduction (nats)
$(2/3, 4/3)$	$+2.54$	-0.40	0.0642
$(1, 1)$	0.00	$+2.15$	0.0887
$(4/3, 2/3)$	-2.20	$+0.94$	0.0717

E.3 Layer Contributions across Tasks

Independent evaluation. We fit probes on each fold's training half and split the opposite half by class. One subset selects regularization, scales, z_0, and the CE-minimizing $\eta \in [0, 1]$; the other evaluates predictions. We include both endpoints when selecting η and exchange the subsets, yielding four evaluations per seed. Averaging these evaluations and six probes within each seed gives $A(s_j; z_0) = 0.2017 \pm 0.0140$ on CREMA-D and 0.5203 ± 0.0082 on KS. The corresponding CE reductions from z_0 are 0.0322 ± 0.0047 and 0.1020 ± 0.0026. Figure 3 instead estimates test set A with OOF-fitted z_0 and scales, averaging the three methods in Table 4 rather than seeds.

Adding and removing individual sources. Using the fixed models and $T/S/E$ split from Appendix E.2, we add one layer predictor to Output only or remove it from +Both, refitting the convex combination on S and evaluating on E. In Table 29, ADD is the accuracy change from adding that source; DROP is +Both accuracy minus accuracy without the source. Thus, negative DROP means removal improves accuracy. Configurations and scales stay fixed.

Audio L4 has higher standalone accuracy than audio L2/L3, yet adding it reduces KS accuracy by 0.444 pp on average. Audio L2/L3 and visual L4 additions improve accuracy in all three seeds on both tasks. Their smaller removal effects indicate that sources can substitute for one another, rather than every layer being independently necessary.

Image–text and trimodal tasks. Tables 30–31 add predictors by modality or individually. Food-101 uses Appendix B.3's six probes with three heads. A separate MOSI layer diagnostic retains four

Table 29: Accuracy changes (pp) from adding or removing a layer predictor, paired across three seeds.

Predictor	CREMA-D ADD	CREMA-D DROP	KS ADD	KS DROP
Audio L2	$+1.255 \pm 0.595$	$+0.050 \pm 0.017$	$+0.505 \pm 0.314$	$+0.003 \pm 0.064$
Audio L3	$+1.255 \pm 0.846$	$+0.241 \pm 0.336$	$+0.559 \pm 0.274$	$+0.086 \pm 0.010$
Audio L4	$+0.030 \pm 0.783$	0.000 ± 0.000	-0.444 ± 0.245	0.000 ± 0.000
Visual L2	$+0.271 \pm 0.533$	0.000 ± 0.000	$+0.192 \pm 0.107$	0.000 ± 0.000
Visual L3	$+0.542 \pm 0.581$	$+0.050 \pm 0.171$	$+0.655 \pm 0.288$	-0.026 ± 0.131
Visual L4	$+0.964 \pm 0.291$	$+0.211 \pm 0.168$	$+1.042 \pm 0.158$	$+0.339 \pm 0.065$

heads and adds nine Transformer predictors at the last valid token of blocks 2, 3, and 5 per modality. Positions are fixed; configurations, scales, and combinations are OOF-selected, with probes refitted on the full model. Test-estimated A is diagnostic only; gains follow combination refitting.

Table 30: Modality-specific additions to Output only: accuracy is addition minus reference; CE reduction reverses this difference.

Dataset	Added predictors	Acc. change (pp)	CE reduction (nats)
Food-101	+Text	0.93 ± 0.05	0.0370 ± 0.0014
	+Image	0.39 ± 0.03	0.0170 ± 0.0011
	+Both	1.08 ± 0.06	0.0413 ± 0.0016
CMU-MOSI	+Text	2.53 ± 3.00	0.0541 ± 0.0305
	+Audio	0.68 ± 0.99	0.0039 ± 0.0044
	+Visual	0.53 ± 1.03	0.0013 ± 0.0015
	+All	2.53 ± 3.00	0.0543 ± 0.0306

Table 31: Single-predictor additions: mean A and paired accuracy/CE gains relative to Output only.

Food-101

Predictor	A	Acc. change	CE reduction
Embedding	0.223	0.45 ± 0.03	0.0172 ± 0.0013
Forward	0.303	0.93 ± 0.05	0.0370 ± 0.0014
Bi-GRU	0.196	0.50 ± 0.03	0.0210 ± 0.0018
Image L2	0.082	0.29 ± 0.06	0.0150 ± 0.0009
Image L3	0.085	0.34 ± 0.04	0.0177 ± 0.0009
Image L4	0.031	0.30 ± 0.04	0.0115 ± 0.0008

CMU-MOSI

Predictor	A	Acc. change	CE reduction
Text 2	0.123	2.48 ± 3.07	0.0550 ± 0.0289
Text 3	0.102	2.28 ± 2.98	0.0431 ± 0.0300
Text 5	0.065	1.60 ± 2.15	0.0335 ± 0.0311
Audio 2	-0.065	0.34 ± 0.51	0.0037 ± 0.0042
Audio 3	-0.064	0.63 ± 0.73	0.0041 ± 0.0051
Audio 5	-0.065	0.63 ± 0.94	0.0041 ± 0.0038
Visual 2	-0.069	0.15 ± 0.58	0.0000 ± 0.0015
Visual 3	-0.065	0.39 ± 1.02	0.0018 ± 0.0017
Visual 5	-0.065	0.39 ± 1.02	0.0018 ± 0.0016

Sharing the models, test samples, and four-head Output only reference with Table 30, predictors fitted to fixed MOSI input features gain 4.81 ± 2.15 pp with the fusion head or 4.86 ± 2.23 when added to all original heads; neither uses Transformer intermediate layers.

E.4 Probe Refitting and Combination Transfer

With OOF scales and weights fixed, probe refitting improves accuracy by 17.74 ± 5.47 pp on CREMA-D and 3.25 ± 0.82 pp on KS over directly applying fold probes and standardizers to full model features and averaging their predictions. Table 32 compares fold models using local probes with full models using refitted probes on the same test examples. Each prediction set retains its

OOF combination, scales, and stacking standardizer. We average fold losses rather than logits, as in Section 3.4.

Table 32: Layer CE gains, Output only minus +Both. Change is full model minus fold model gain.

Dataset	Combination	Fold models	Full model	Change
CREMA-D	Convex	0.0670 ± 0.0028	0.0478 ± 0.0116	-0.0192 ± 0.0089
CREMA-D	Linear stacking	0.0375 ± 0.0042	0.0171 ± 0.0083	-0.0204 ± 0.0122
Kinetics-Sounds	Convex	0.1599 ± 0.0062	0.1083 ± 0.0061	-0.0516 ± 0.0089
Kinetics-Sounds	Linear stacking	0.1254 ± 0.0042	0.0856 ± 0.0101	-0.0398 ± 0.0072

Against the test-fitted three-head simplex optimum, convex LAMF reduces CE by 0.0453 ± 0.0097 nats on CREMA-D and 0.1031 ± 0.0080 nats on KS, with positive reductions in every seed. These comparisons do not bound Γ or certify Proposition 3's sufficient condition.