DECOUPLING IS NOT ENOUGH: UNDERSTANDING GENERALIZATION IN MULTIMODAL FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modality imbalance often prevents multimodal learning from fully exploiting the information available in each modality. Although decoupling unimodal learning from fusion alleviates optimization interference, it does not ensure fusion generalization. We identify a remaining source of failure: fusion trained on samples used for unimodal fitting can exploit modality reliability and error patterns that transfer poorly to unseen data. We analyze this failure mechanism and propose a generalization-aware multimodal fusion framework. The framework constructs out-of-fold representations and predictions to learn feature-based corrections to averaged unimodal logits from out-of-sample evidence. To account for differences among the unimodal models fitted across folds, we formulate a cross-fold objective that balances average correction gains against unfavorable fold-specific effects. Experiments across multimodal benchmarks support the effectiveness of the framework, with component ablations demonstrating the complementary benefits of out-of-fold fusion learning and cross-fold regularization.

1 INTRODUCTION

Multimodal learning integrates heterogeneous sources of information to improve predictive performance. However, joint training can suffer from modality imbalance (Wang et al., 2020; Peng et al., 2022; Huang et al., 2022), where optimization disproportionately favors certain modalities and leaves others insufficiently learned, limiting the benefits of multimodal integration. Recent approaches address this issue through modality-aware optimization, auxiliary learning objectives, data-level interventions, and alternative learning paradigms. Among these, decoupled optimization (Wei et al., 2025; Feng et al., 2026) seeks to alleviate interference between unimodal representation learning and multimodal fusion. This raises a further question: is removing such optimization interference sufficient for learned fusion to improve upon a simple combination of independently trained unimodal predictions?

Taking CREMA-D as a case study, we compare representative recent multimodal learning methods with independently trained unimodal models whose predictions are combined by simply averaging their logits, without learning any additional fusion parameters. As shown in Fig. 1(a), the evaluated methods fall short of this simple baseline. To investigate a possible explanation, we additionally examine the training-set performance of the independent unimodal models. Although video achieves higher accuracy than audio on the test set, audio fits the training set substantially better than video, reversing their relative accuracy across the two settings. This observation suggests that modality reliability inferred from training performance may not reflect reliability on unseen data. We therefore hypothesize that, even when imbalanced representation learning is alleviated, fusion learning may remain vulnerable to this discrepancy: modality relationships learned from training samples may transfer poorly when relative reliability changes substantially or reverses on unseen samples, potentially yielding worse performance than simple logit averaging.

To investigate this hypothesis, we decouple unimodal training from fusion learning to prevent fusion optimization from affecting the unimodal representations. We independently train the unimodal models on 90% of the original training set and then freeze them. The remaining 10%, unseen during unimodal training, is used to train the fusion module. As a control, we train the same fusion architecture on an equally sized subset drawn from the unimodal training data, using identical pretrained unimodal models and fusion initialization in both settings. As shown in Fig. 1(b), fusion trained

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.

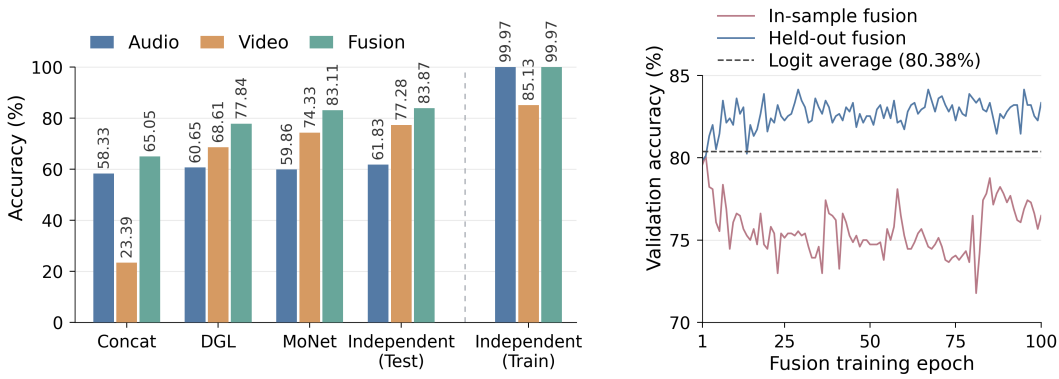


Figure 1: Motivating observations on CREMA-D. **Left**: Test accuracies of the evaluated methods and independently trained unimodal models, alongside the training accuracies of the latter. For the independent models, fusion consists of parameter-free logit averaging. **Right**: Validation accuracy when the same fusion architecture is trained on equally sized subsets seen or unseen during unimodal training, using identical fixed unimodal predictors. The dashed line denotes their logit-averaging baseline.

on seen samples reduces its fitting loss but degrades validation accuracy relative to logit averaging, whereas fusion trained on unseen samples yields positive validation gains over most of training. These results show that decoupling alone is insufficient: even with fixed unimodal representations, fusion can learn corrections that reduce training loss while harming generalization.

Motivated by these observations, we analyze how seen-sample fusion can reinforce unimodal decision rules that generalize poorly, and propose an out-of-fold residual fusion framework. For each training sample, representations and predictions are generated by independently trained unimodal models fitted without that sample. A shared fusion head learns feature-based corrections to averaged unimodal logits from these out-of-sample representations and predictions, allowing all training samples to contribute to fusion learning. To account for differences among the fitted unimodal models, we further introduce cross-fold robust regularization that balances the average loss change relative to logit averaging against the least favorable fold, discouraging corrections that improve average performance at the expense of particular predictor sets. Our contributions are summarized as follows:

- We identify a discrepancy between in-sample and out-of-sample unimodal behavior that can undermine fusion generalization even under decoupled optimization, and analyze how seen-sample fusion can reinforce unimodal decision rules that transfer poorly to unseen data.
- We propose an out-of-fold residual fusion framework that learns corrections from out-of-sample unimodal representations and predictions, together with cross-fold robust regularization that balances average loss changes against unfavorable fold-specific effects relative to logit averaging.
- We evaluate the framework across multimodal benchmarks against recent methods and independent unimodal ensembles. Component ablations demonstrate the complementary benefits of out-of-fold fusion learning and cross-fold regularization, while experiments with different fusion heads establish compatibility with multiple fusion architectures.

2 Related Work

2.1 Multimodal Imbalanced Learning

Joint multimodal training can exhibit modality competition (Huang et al., 2022) and persistent unimodal bias, leaving some modalities insufficiently learned (Huang et al., 2022; Wu et al., 2022). Modality-aware optimization addresses this imbalance through gradient modulation (Peng et al., 2022; Li et al., 2023; Wei & Hu, 2024; Hu et al., 2025; Jiang et al., 2025) and flatness-oriented optimization (Chen et al., 2026; Rajoli Nowdeh et al., 2026; Wang et al., 2026). Auxiliary learning

objectives promote unimodal feature acquisition through prototype guidance (Fan et al., 2023), unimodal distillation (Du et al., 2023), and information regularization (Huang et al., 2025). Data-level interventions improve learning balance through sample-level modality valuation (Wei et al., 2024), data remixing (Ma et al., 2025), and modality-misaligned augmentation (Hwang et al., 2026). Alternative learning paradigms restructure training through alternating adaptation (Zhang et al., 2024), boosting (Hua et al., 2024), gradient disentanglement (Wei et al., 2025), and nested optimization (Feng et al., 2026). These approaches improve modality learning and its coordination with fusion. We investigate a limitation that remains after optimization interference is removed: even with independently trained and fixed unimodal predictors, fusion learned from their in-sample behavior can underperform simple logit averaging.

2.2 Robust Multimodal Fusion

Prior work links multimodal generalization failures to differences in modality-specific overfitting (Wang et al., 2020) and regularizes complementary and contradictory features to improve fusion (Tejero-de Pablos, 2024). Information-theoretic approaches preserve task-relevant information through mutual information maximization (Han et al., 2021) and characterize modality interactions through redundancy, uniqueness, and synergy (Liang et al., 2023). These interactions are further modeled using specialized experts (Xin et al., 2025) and sample-specific information decomposition (Yang et al., 2026). Related objectives balance modality contributions through game-theoretic regularization (Kontras et al., 2026b) or encourage cross-modal synergy through information bottlenecks (Kontras et al., 2026a). Under low-quality data conditions (Zhang et al., 2026), quality-aware fusion dynamically weights modality predictions (Zhang et al., 2023), while conformal calibration adjusts representations and gradient updates according to estimated reliability (Jiang et al., 2026). We investigate a complementary source of fusion failure: participation in unimodal training can alter the reliability and error patterns presented to fusion, causing relationships learned on seen samples to transfer poorly to unseen data even after optimization is decoupled. Our framework addresses this discrepancy through out-of-fold fusion training and regularizes correction gains across fitted unimodal predictors.

3 Preliminaries and Problem Analysis

3.1 Problem Setup and Notation

We consider C-class classification with M modalities and input–label pairs $(\mathbf{x}, y) \sim P$, where $\mathbf{x} = (x^{(1)}, \ldots, x^{(M)})$ and $y \in \{1, \ldots, C\}$. For modality m, an encoder h_m produces a representation $\mathbf{z}^{(m)}$, and a linear head c_m computes the unimodal logits:

$$\mathbf{z}^{(m)} = h_m(x^{(m)}), \qquad \mathbf{s}^{(m)} = c_m(\mathbf{z}^{(m)}) = A_m \mathbf{z}^{(m)} + \mathbf{a}_m. \tag{1}$$

Let ϕ_m collect the parameters of h_m and c_m. For the mechanistic analysis below, we consider a linear fusion head:

$$\mathbf{s}_\theta = g_\theta([\mathbf{z}^{(1)}; \ldots; \mathbf{z}^{(M)}]) = \sum_{m=1}^{M} W_m \mathbf{z}^{(m)} + \mathbf{b}, \tag{2}$$

where $[\cdot;\cdot]$ denotes concatenation and $\theta = (W_1, \ldots, W_M, \mathbf{b})$. A fixed average of unimodal logits plus an affine residual also has this form; their equivalence is detailed in Appendix B.1. To separate unimodal learning from fusion, we consider decoupled gradient updates, following the principle of DGL (Wei et al., 2025). The corresponding objectives are

$$\mathcal{L}_m = \mathbb{E}_{\mathcal{D}_{\text{uni}}}\left[\ell(\mathbf{s}^{(m)}, y)\right], \qquad \mathcal{L}_{\text{fus}} = \mathbb{E}_{\mathcal{D}_{\text{fus}}}[\ell(\mathbf{s}_\theta, y)], \tag{3}$$

where ℓ is cross-entropy and $\mathbb{E}_{\mathcal{D}}$ denotes the empirical average over dataset $\mathcal{D}$. Each objective updates only its associated parameters:

$$\phi_m \leftarrow \phi_m - \eta_{\text{uni}} \nabla_{\phi_m} \mathcal{L}_m, \qquad \theta \leftarrow \theta - \eta_{\text{fus}} \nabla_\theta \mathcal{L}_{\text{fus}}, \tag{4}$$

with learning rates η_{uni} and η_{fus}. This specifies the gradient paths without imposing a training schedule. Fusion samples are termed *seen* if used to fit the unimodal models and *unseen* if excluded from that fitting. Although decoupling blocks fusion gradients from updating the unimodal models, our introductory experiments (Fig. 1) show that seen-sample fusion can still degrade generalization. We next analyze this failure with the unimodal models held fixed.

3.2 Generalization Failure in Seen-Sample Fusion

Decoupled optimization does not prevent fusion from reinforcing unimodal decision rules that generalize poorly. To examine this mechanism, we fix the unimodal models and train fusion on seen samples. The linear classifier c_m induces a direction in the fusion parameter space:

$$D_m = (0, \ldots, \underbrace{A_m}_{W_m \text{ block}}, \ldots, 0, \mathbf{a}_m), \tag{5}$$

where $\mathbf{a}_m$ occupies the fusion bias block. For all parameter projections, we use class-centered weights and biases, which preserve softmax probabilities and classification margins (Appendix B.1). A displacement along D_m adds a scaled copy of the corresponding unimodal logits to the fused prediction. To determine whether fusion optimization favors this direction, we examine its alignment with the negative fusion gradient:

$$-\langle \nabla_\theta \mathcal{L}_{\text{fus}}, D_m \rangle = \mathbb{E}_{\mathcal{D}_{\text{fus}}} \left[\sum_{k \neq y} p_{\theta,k}(\mathbf{x}) \big(s_y^{(m)}(\mathbf{x}) - s_k^{(m)}(\mathbf{x}) \big) \right], \tag{6}$$

where $p_{\theta,k}$ is the kth softmax probability of the fused logits and the inner product sums over all parameter entries. A positive value indicates that an ordinary gradient descent step increases the parameter component along D_m. For a fixed direction and learning rate, a larger value gives a larger increase. The alignment is positive when the modality strictly separates the seen samples, or more generally when fusion-weighted positive margins outweigh negative ones (Appendix B.2). Unimodal fitting can make this condition hold on seen samples, causing fusion optimization to repeatedly reinforce the fitted decision rule. On unseen samples, unimodal errors introduce negative margins that can oppose this reinforcement. Training fusion on seen samples may therefore favor a direction that strengthens incorrect unimodal predictions on unseen inputs.

To examine actual optimization dynamics, we measure $q_m(t) = \langle \theta_t - \theta_0, D_m / \|D_m\| \rangle$ on CREMA-D, where θ_0 is the saved fusion initialization and θ_t denotes effective fusion parameters. These projections use recorded parameter changes and therefore include the effects of the optimizer. The upper panel of Fig. 2 shows substantially larger cumulative projections along the audio direction than the video direction. This observation is consistent with Fig. 1(a), where audio achieves nearly perfect training accuracy while video retains substantial error; accuracy alone does not establish the ordering of the projections. We then remove $q_a(t) D_a / \|D_a\|$ at each checkpoint while preserving the initialization and the parameter change orthogonal to D_a. At the final seen checkpoint, removal restores 28 of the 43 validation samples that fusion changes from correct to incorrect relative to logit averaging. Across all validation samples, it repairs 38 errors and introduces 7, raising accuracy from 75.13% to 79.30% (lower panel of Fig. 2). The exact intervention and numerical accounting appear in Appendices B.3 and B.4. These results link the accumulated audio-direction component to the observed generalization failure.

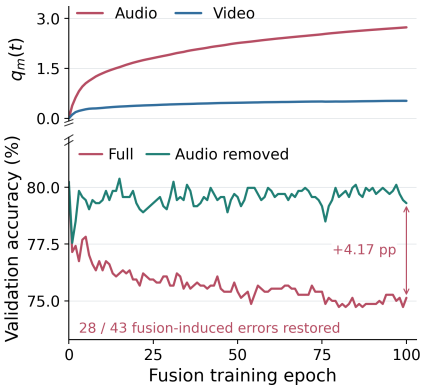


Figure 2: Seen-sample fusion on CREMA-D. Top: cumulative projections onto the class-centered, normalized audio and video classifier directions. Bottom: validation accuracy before and after removing the accumulated audio-direction component.

3.3 Why Learn Fusion from Unseen Samples?

Training fusion on unseen samples exposes it to unimodal behavior beyond the data used for unimodal fitting. Let $U = \{(\mathbf{x}_i, y_i)\}_{i=1}^q$ contain i.i.d. samples from P, independent of the data and randomness used to fit and select the unimodal models. With the fitted parameters $\widehat{\phi}$ held fixed, define

$$\widehat{R}_U(\theta) = \frac{1}{q} \sum_{i=1}^{q} \ell(\mathbf{s}_\theta(\mathbf{x}_i), y_i), \qquad R(\theta) = \mathbb{E}_{(\mathbf{x},y)\sim P}[\ell(\mathbf{s}_\theta(\mathbf{x}), y)]\,. \tag{7}$$

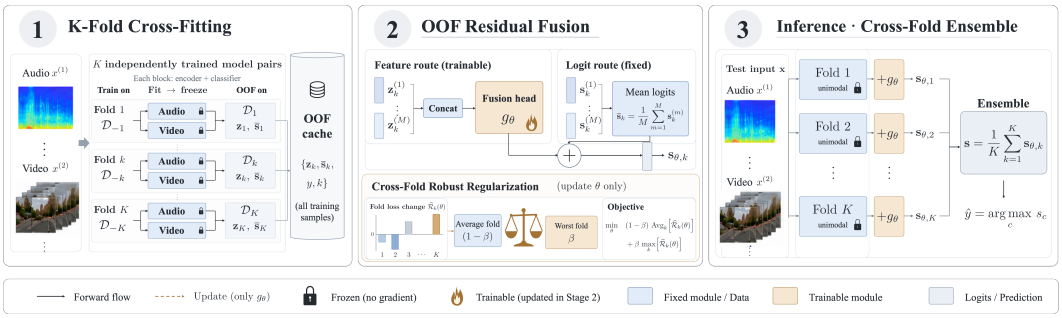


Figure 3: Overview of the proposed framework. Unimodal models are trained independently on $K-1$ folds to generate representations and logits for the held-out fold. A shared fusion head learns corrections to averaged unimodal logits from the pooled OOF records. Cross-fold robust regularization balances average improvement against unfavorable fold-specific loss changes. At inference, corrected logits are averaged across folds.

Here, $\widehat{R}_U$ is the empirical fusion risk on U, and R is the population risk of the same fitted predictor on independent samples; dependence on $\widehat{\phi}$ is suppressed. For each fixed fusion head θ, integrability and independence yield

$$\mathbb{E}_U\left[\widehat{R}_U(\theta) \mid \widehat{\phi}\right] = R(\theta), \qquad \widehat{R}_U(\theta) \xrightarrow[q\to\infty]{\text{a.s.}} R(\theta). \tag{8}$$

The proof is given in Appendix B.5. This identity is not generally guaranteed for seen samples, which also determine the fitted unimodal parameters. These properties motivate larger independent unseen sets for fusion learning, without guaranteeing the generalization of a head optimized on U. Since enlarging a single holdout reduces the data available for unimodal training, we construct OOF representations and predictions, allowing every training sample to supervise fusion while remaining excluded from the fitting of its corresponding unimodal models. This construction follows the sample-exclusion principle; the i.i.d. statement above does not automatically extend to pooled OOF records or to a head fitted on them (Appendix B.6).

4 Method

As illustrated in Fig. 3, this construction separates unimodal learning from fusion while allowing every training sample to supervise corrections through out-of-sample unimodal behavior. Since these corrections are shared across differently fitted predictors, we further regularize their fold-specific effects to discourage gains that come at the expense of particular predictor sets.

4.1 Out-of-Fold Residual Fusion

Based on the preceding analysis, we construct out-of-fold (OOF) representations and predictions to expand unseen supervision without reserving a larger holdout. We partition $\mathcal{D}$ into K disjoint, approximately class-stratified folds $\mathcal{D}_1, \ldots, \mathcal{D}_K$. For each fold k, we independently train each modality on $\mathcal{D}_{-k} = \mathcal{D} \setminus \mathcal{D}_k$:

$$\min_{h_m, c_m} \frac{1}{|\mathcal{D}_{-k}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_{-k}} \ell\Big(c_m(h_m(x^{(m)})), y\Big), \qquad m = 1, \ldots, M. \tag{9}$$

We select each encoder and its classifier jointly according to unimodal validation accuracy and freeze the resulting models, denoted by $\widehat{h}_{m,-k}$ and $\widehat{c}_{m,-k}$. For each sample in $\mathcal{D}_k$, these models produce

$$\mathbf{z}_k^{(m)}(\mathbf{x}) = \widehat{h}_{m,-k}(x^{(m)}), \qquad \mathbf{s}_k^{(m)}(\mathbf{x}) = \widehat{c}_{m,-k}\Big(\mathbf{z}_k^{(m)}(\mathbf{x})\Big). \tag{10}$$

We concatenate the representations as $\mathbf{z}_k = [\mathbf{z}_k^{(1)}; \ldots; \mathbf{z}_k^{(M)}]$ and average the unimodal logits:

$$\bar{\mathbf{s}}_k(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{s}_k^{(m)}(\mathbf{x}). \tag{11}$$

Pooling these representations and predictions across folds allows every training sample to supervise fusion while remaining unseen during the fitting of its corresponding unimodal models. A fusion head g_θ, shared across folds, learns a feature-based correction to the averaged logits:

$$\mathbf{s}_{\theta,k}(\mathbf{x}) = \bar{\mathbf{s}}_k(\mathbf{x}) + g_\theta(\mathbf{z}_k(\mathbf{x})). \tag{12}$$

With all unimodal models fixed, the basic fusion objective minimizes the average cross-entropy across folds:

$$\mathcal{L}_{\mathrm{OOF}}(\theta) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\mathcal{D}_k|}\sum_{(\mathbf{x},y)\in\mathcal{D}_k}\ell(\mathbf{s}_{\theta,k}(\mathbf{x}), y)\,. \tag{13}$$

4.2 Cross-Fold Robust Regularization

Although OOF construction exposes fusion to unseen unimodal behavior, fold-specific performance varies with both the fitted unimodal predictors and the held-out samples. A shared correction can therefore yield uneven gains across folds. We account for this variability by measuring the change in each fold's loss relative to unimodal logit averaging:

$$\widehat{\mathcal{R}}_k(\theta) = \frac{1}{|\mathcal{D}_k|}\sum_{(\mathbf{x},y)\in\mathcal{D}_k}\left[\ell(\mathbf{s}_{\theta,k}(\mathbf{x}), y) - \ell(\bar{\mathbf{s}}_k(\mathbf{x}), y)\right]. \tag{14}$$

Negative values indicate that the learned correction improves upon logit averaging. We combine the average loss change with the least favorable fold:

$$\min_\theta\ \widehat{\Phi}_\beta(\theta) = \frac{1-\beta}{K}\sum_{k=1}^{K}\widehat{\mathcal{R}}_k(\theta) + \beta\max_{1\le k\le K}\widehat{\mathcal{R}}_k(\theta). \tag{15}$$

where $\beta \in [0,1]$ controls the emphasis on the least favorable fold. The objective balances the average loss change against the least favorable fold, discouraging fold-specific degradation without requiring identical gains across folds. At $\beta = 0$, the objective differs from $\mathcal{L}_{\mathrm{OOF}}$ only by a constant; increasing β places greater emphasis on the fold with the largest loss increase or smallest loss reduction.

At inference, we average the corrected logits across folds:

$$\mathbf{s}_{\mathrm{ens},\theta}(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K}\left[\bar{\mathbf{s}}_k(\mathbf{x}) + g_\theta(\mathbf{z}_k(\mathbf{x}))\right]. \tag{16}$$

5 Experiments

5.1 Experiment Setup

We evaluate our framework on seven benchmarks spanning three multimodal settings. For audio and visual classification, we use CREMA-D (Cao et al., 2014), AVE (Tian et al., 2018), and Kinetics-Sounds (Arandjelovic & Zisserman, 2017), with separate ResNet-18 (He et al., 2016) encoders for audio spectrograms and video frames. For image and text classification, we use MVSA-Single (Niu et al., 2016) and UPMC-Food101 (Wang et al., 2015), employing ResNet-18 for images and a bidirectional GRU (Chung et al., 2014) for text. For trimodal learning, we evaluate humor recognition on UR-FUNNY (Hasan et al., 2019) and sentiment classification on CMU-MOSEI (Zadeh et al., 2018), both using language, audio, and visual inputs. UR-FUNNY employs separate Transformer (Vaswani et al., 2017) encoders for language, audio, and visual features, whereas CMU-MOSEI uses independent unidirectional GRU encoders, followed by linear projections of their final hidden states to obtain modality-specific representations.

Our baselines include Summation and Concatenation, which jointly optimize modality encoders with additive and concatenation-based fusion, respectively, and UniAvg, which averages the logits of independently trained unimodal models. We further compare against nine representative multimodal learning methods: OGM-GE (Peng et al., 2022), AGM (Li et al., 2023), MMPareto (Wei & Hu, 2024), MLA (Zhang et al., 2024), DGL (Wei et al., 2025), InfoReg (Huang et al., 2025),

Table 1: Classification accuracy (%) on five multimodal benchmarks. Multi denotes multimodal performance. Results for CREMA-D, AVE, MVSA, and UPMC-Food101 are taken from MoNet (Feng et al., 2026), with standard deviations omitted. Unfilled entries are indicated by –.

Methods	CREMA-D			AVE			Kinetics-Sounds			MVSA			UPMC-Food101		
	Audio	Visual	**Multi**	Audio	Visual	**Multi**	Audio	Visual	**Multi**	Image	Text	**Multi**	Image	Text	**Multi**
Summation	53.15	22.18	61.07	53.73	19.32	66.33	35.52	35.82	57.38	28.45	65.64	72.19	29.42	82.40	87.58
Concatenation	54.75	24.51	63.26	53.48	18.33	66.42	37.48	34.35	58.30	29.67	61.72	71.16	30.11	87.50	87.50
UniAvg	61.83	77.28	83.87	63.59	39.14	71.07	41.21	63.01	73.06	52.12	69.69	69.69	54.64	85.02	90.72
OGM-GE	57.26	64.92	76.43	50.66	19.82	65.67	35.27	42.99	60.75	29.61	60.18	70.84	30.39	81.66	87.49
AGM	57.66	26.84	65.68	54.23	17.49	68.32	37.35	39.56	59.46	30.12	42.83	67.63	30.30	79.57	87.18
MMPareto	61.02	46.64	69.53	61.36	33.17	69.73	42.01	63.04	73.36	50.05	74.63	72.04	49.60	84.06	89.24
MLA	59.14	65.05	74.42	64.10	32.18	68.74	42.13	58.60	72.08	51.57	68.80	61.87	51.70	83.53	89.30
DGL	61.88	72.96	79.40	63.60	24.87	68.82	42.89	62.23	73.19	50.80	51.32	69.62	56.51	67.51	88.24
InfoReg	59.13	61.00	74.77	63.25	31.44	69.89	43.31	60.32	72.81	50.67	74.57	74.05	40.37	84.26	89.20
MASAM	60.54	76.23	82.91	63.34	38.78	70.87	42.07	62.77	74.11	53.76	73.37	74.92	53.64	84.79	91.47
TCNE	61.03	75.60	83.27	63.32	38.21	70.24	41.21	63.04	74.56	52.68	73.13	74.13	54.46	84.78	91.32
MoNet	59.86	74.33	83.11	63.10	39.39	71.23	42.31	63.78	75.74	50.93	72.19	74.25	54.14	84.91	90.42
Ours	64.52	80.38	**89.24**	65.17	39.05	**71.64**	47.27	64.30	**77.34**	52.22	75.11	**75.53**	62.04	85.11	**92.04**

Table 2: Classification accuracy (%) on two trimodal benchmarks. Multi denotes performance using all three modalities.

Dataset		Summation	Concatenation	UniAvg	OGM-GE	AGM	MM Pareto	MLA	DGL	InfoReg	MASAM	TCNE	MoNet	**Ours**
CMU-MOSEI	Audio	62.88	37.45	64.23	37.12	62.30	63.02	62.99	37.15	62.91	63.02	54.24	62.88	63.59
	Visual	57.84	53.85	64.94	57.68	57.95	62.93	63.98	37.15	62.63	64.42	64.58	63.87	63.62
	Text	81.04	82.23	82.17	82.87	81.99	82.90	82.90	82.62	80.97	82.98	82.68	82.65	83.01
	Multi	81.76	82.39	80.15	83.09	82.70	82.73	83.01	82.76	82.37	83.03	<u>83.23</u>	82.84	**83.79**
UR-FUNNY	Audio	50.91	50.60	52.72	55.03	52.01	54.12	53.02	55.33	54.63	53.12	50.60	54.33	54.63
	Visual	49.30	49.50	58.05	54.93	54.83	59.46	53.82	50.70	57.24	57.65	58.25	59.66	57.24
	Text	58.25	58.85	59.36	59.36	56.54	55.63	58.45	59.26	58.95	62.88	59.15	61.77	62.27
	Multi	59.15	60.56	58.85	58.55	58.35	58.45	59.56	59.05	61.97	62.17	61.17	<u>62.47</u>	**67.20**

MASAM (Chen et al., 2026), TCNE (Yu et al., 2026), and MoNet (Feng et al., 2026). We report classification accuracy and select checkpoints using validation accuracy, with test results reported for the selected checkpoints. Additional dataset descriptions, preprocessing procedures, and training configurations are provided in Appendix A.

5.2 Main Result

Results on bimodal benchmarks. Table 1 shows that our framework achieves the highest multimodal accuracy across all five benchmarks. It improves upon the best competing multimodal result by 5.37, 0.41, 1.60, 0.61, and 0.57 percentage points on CREMA-D, AVE, Kinetics-Sounds, MVSA, and UPMC-Food101, respectively. Notably, UniAvg already outperforms all nine recent methods on CREMA-D, yet our framework further raises accuracy from 83.87% to 89.24%. Across all five datasets, our fused predictions outperform both UniAvg and the strongest unimodal branch within our framework, supporting the effectiveness of learned fusion beyond independent unimodal prediction and simple logit averaging.

Results on trimodal benchmarks. As shown in Table 2, our method achieves the highest multimodal accuracy on both datasets, reaching 83.79% on CMU-MOSEI and 67.20% on UR-FUNNY. It outperforms the strongest competing methods, TCNE and MoNet, by 0.56 and 4.73 percentage points, respectively, and improves upon UniAvg by 3.64 and 8.35 percentage points. These gains occur despite substantial differences in unimodal accuracy, particularly on CMU-MOSEI, where text is considerably stronger than audio and vision. On both datasets, our fused predictions also outperform the strongest unimodal branch within our framework, supporting its effectiveness in combining three modalities with unequal predictive capabilities.

5.3 Further Analysis

Component ablation. Table 3 examines the contributions of independent unimodal logit averaging (UA), residual fusion (RF), out-of-fold training (K-fold), and robust regularization (RR). Adding RF

Table 3: Component ablation.

UA	RF	K-fold	RR	CREMA-D	AVE	MVSA
✓				83.87	70.89	70.54
✓	✓			77.55	69.39	67.12
✓		✓		84.54	71.14	69.56
✓	✓	✓		88.98	71.39	72.45
✓	✓	✓	✓	**89.24**	**71.64**	**75.53**

Table 4: Fusion-head ablation.

Fusion head	CREMA-D	AVE	MVSA
UniAvg	84.54	71.14	69.56
Linear	88.98	70.90	73.60
MLP	89.24	71.64	75.53
FiLM	88.98	71.64	74.18
Gated	89.65	71.14	75.53

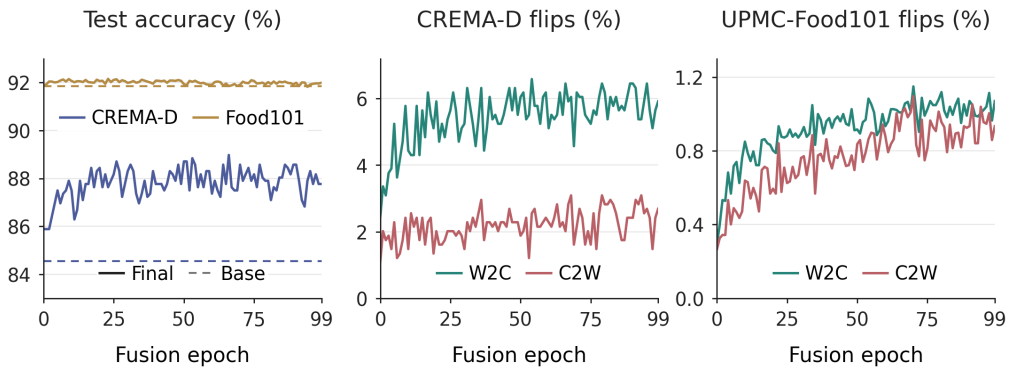


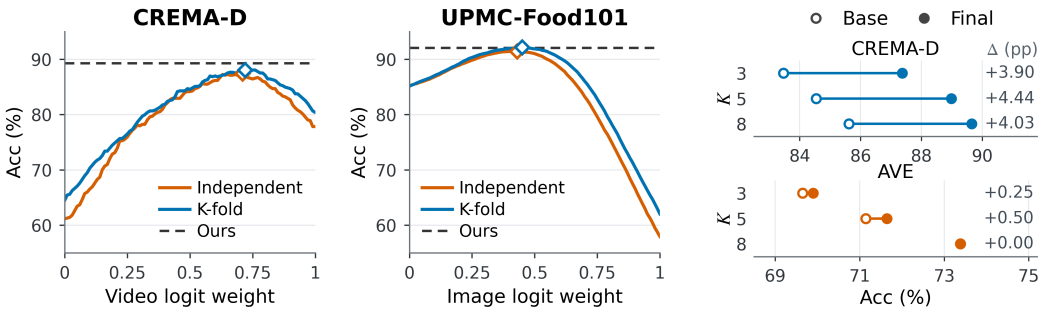
Figure 4: Residual fusion dynamics on CREMA-D and UPMC-Food101 over 100 training epochs. Performance is evaluated relative to UniAvg. Correction and degradation rates denote the fractions of UniAvg's incorrect and correct predictions that fusion changes to correct and incorrect, respectively.

trained on seen samples to UA reduces accuracy on all three datasets, indicating that additional fusion capacity alone does not ensure improvement. K-fold ensembling without RF yields only modest gains on CREMA-D and AVE and decreases accuracy on MVSA. This configuration uses the same fold-specific unimodal models as our full framework, allowing us to distinguish the contribution of learned fusion from that of ensembling. In contrast, learning RF from out-of-fold representations improves accuracy over the corresponding K-fold ensemble by 4.44, 0.25, and 2.89 percentage points on CREMA-D, AVE, and MVSA, respectively. This comparison shows that the gains extend beyond ensembling multiple sets of unimodal models and supports learning corrections from unseen unimodal behavior. RR further improves accuracy on all three datasets, with the largest gain of 3.08 percentage points on MVSA, supporting the effectiveness of the cross-fold objective.

Compatibility with different fusion heads. Table 4 compares Linear, MLP, FiLM, and Gated residual fusion heads while keeping the pretrained unimodal models, OOF data, and robust regularization settings fixed. All four heads outperform UniAvg on CREMA-D and MVSA, with gains ranging from 4.44 to 5.11 and from 4.04 to 5.97 percentage points, respectively. Improvements on AVE are more modest: MLP and FiLM improve accuracy by 0.50 percentage points, while Gated matches UniAvg and Linear performs slightly worse. These results demonstrate compatibility with several fusion architectures, although the benefit depends on the dataset and fusion head.

Residual fusion dynamics and error correction. Figure 4 tracks test accuracy and prediction changes over 100 fusion-training epochs to assess the consistency of residual fusion gains throughout training. Relative to UniAvg, fusion improves accuracy at every epoch on CREMA-D, with a mean gain of 3.26 percentage points. On UPMC-Food101, gains are positive in 97 of 100 epochs but smaller in magnitude, averaging 0.15 percentage points. Averaged across epochs, fusion corrects 35.3% of UniAvg's errors on CREMA-D while making 2.6% of its correct predictions incorrect; the corresponding rates on UPMC-Food101 are 11.3% and 0.84%. These trajectories show that corrected errors generally outnumber newly introduced errors, yielding sustained, though dataset-dependent, accuracy gains.

Comparison with global modality reweighting. Figure 5(a) examines whether the gains of residual fusion can be reproduced by global modality weighting. We sweep $\alpha \in [0, 1]$ for the weighted logits



(a) Comparison with global modality reweighting (b) K-fold ablation

Figure 5: Comparison with global modality reweighting and sensitivity to the number of folds. **(a)** Global logit-weight sweeps for independently trained unimodal models and their K-fold ensembles, compared with our method. **(b)** Accuracy of UniAvg and our method under different numbers of folds on CREMA-D and AVE.

$\alpha \mathbf{s}^{(1)} + (1-\alpha)\mathbf{s}^{(2)}$, using both independently trained unimodal models and K-fold ensembles. For the latter, predictions are first averaged across folds within each modality. On CREMA-D, the highest accuracies in the sweep are 87.37% and 88.04%, respectively, both below our method's 89.24%. On UPMC-Food101, validation-selected weights yield 91.38% and 92.10%, compared with 92.04% for our method. These results indicate that residual fusion provides gains beyond global modality reweighting on CREMA-D, whereas reweighting the K-fold ensemble is already competitive on UPMC-Food101.

Sensitivity to the number of folds. Figure 5(b) evaluates the effect of K on CREMA-D and AVE. Increasing K from 3 to 5 improves accuracy by 1.61 and 1.74 percentage points, respectively, with $K = 5$ yielding the largest gains over the corresponding UniAvg predictors among the tested settings. At $K = 8$, absolute accuracy increases, but the gains from residual fusion decline to 4.03 and 0.00 percentage points. The increase in absolute accuracy therefore reflects improvements in the underlying unimodal ensemble, whose constituent models are trained on larger subsets as K increases. Since $K = 8$ also requires training three additional sets of unimodal models, we use $K = 5$ to balance predictive performance and computational cost.

6 Conclusion

In this work, we identify training membership as a source of generalization failure in decoupled multimodal fusion. Our analysis and controlled experiments show how fusion trained on seen samples can reinforce unimodal decision rules that amplify errors on unseen data. We address this issue through out-of-fold residual fusion, which learns corrections to unimodal logit averaging from out-of-sample representations and predictions. Cross-fold robust regularization further discourages corrections with unfavorable fold-specific effects. Experiments across seven benchmarks support the framework's effectiveness, while ablations distinguish the contributions of learned corrections, unimodal ensembling, and regularization. These findings highlight that effective fusion depends on both the construction of its training data and the transferability of its corrections across fitted unimodal predictors.

AI use statement

During the preparation of this manuscript, generative AI tools assisted with language refinement, literature searches, experimental programming, figure creation, and the identification of possible errors or inconsistencies in mathematical derivations. All material developed with this assistance underwent critical review by the authors. Bibliographic information and related claims were cross-checked against the original publications, experimental and visualization code was examined against the underlying data, and mathematical derivations were independently assessed. Responsibility for the final work rests entirely with the authors and extends to all text, code, figures, results, and claims, including those developed with generative AI assistance.

REFERENCES

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617. IEEE, 2017.

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.

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*, volume 2026, pp. 92384–92411, 2026.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pp. 8632–8656. PMLR, 2023.

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

Yanglin Feng, Yang Qin, Dezhong Peng, Rui Wang, Xiaomin Song, and Peng Hu. Multimodal nested learning for decoupled and coordinated optimization. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=8XOncDWc5j.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 9180–9192, 2021.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2046–2056, 2019.

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.

Disen Hu, Xun Jiang, Zhe Sun, Hao Yang, Chong Peng, Peng Yan, Heng Tao Shen, and Xing Xu. Geometric gradient divergence modulation for imbalanced multimodal learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1337–1345, 2025.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=93gjGDwqim.

Chengxiang Huang, Yake Wei, Zequn Yang, and Di Hu. Adaptive unimodal regulation for balanced multimodal information acquisition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25854–25863. IEEE, 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 *International conference on machine learning*, pp. 9226–9259. PMLR, 2022.

Seong-Hyeon Hwang, Soyoung Choi, and Steven Whang. Midas: Misalignment-based data augmentation strategy for imbalanced multimodal learning. *Advances in Neural Information Processing Systems*, 38:147773–147795, 2026.

Qing-Yuan Jiang, Zhouyang Chi, and Yang Yang. Interactive multimodal learning via flat gradient modification. In *IJCAI*, pp. 5489–5497, 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. *arXiv preprint arXiv:2605.03820*, 2026.

Konstantinos Kontras, Teodora Gagaleska, Thomas Strypsteen, Christos Chatzichristos, Matthew Blaschko, Maarten De Vos, and Paul Pu Liang. Synib: Informational bottleneck for maximizing synergy in multimodal learning. *arXiv preprint arXiv:2606.09853*, 2026a.

Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul Liang, Matthew Blaschko, and Maarten De Vos. Balancing multimodal training through game-theoretic regularization. *Advances in Neural Information Processing Systems*, 38:167953–167982, 2026b.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22157–22167. IEEE, 2023.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023.

Xiaoyu Ma, Hao Chen, and Yongjian Deng. Improving multimodal learning balance and sufficiency through data remixing. *arXiv preprint arXiv:2506.11550*, 2025.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *International conference on multimedia modeling*, pp. 15–27. Springer, 2016.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8237. IEEE, 2022.

Hossein Rajoli Nowdeh, Jie Ji, Xiaolong Ma, and Fatemeh Afghah. Modality-aware sam: Sharpness-aware-minimization driven gradient modulation for harmonized multimodal learning. *Advances in Neural Information Processing Systems*, 38:168772–168794, 2026.

Antonio Tejero-de Pablos. Complementary-contradictory feature regularization against multimodal overfitting. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5667–5676. IEEE, 2024.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European conference on computer vision*, pp. 252–268. Springer, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12692–12702. IEEE, 2020.

Xiaohan Wang, Zhangtao Cheng, Ting Zhong, Leiting Chen, and Fan Zhou. Modality-balanced collaborative distillation for multi-modal domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26535–26543, 2026.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International conference on multimedia & expo workshops (ICMEW)*, pp. 1–6. IEEE, 2015.

Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*, 2024.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27328–27337. IEEE, 2024.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L Ballard, Tianlong Chen, and Qi Long. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. *Proceedings of machine learning research*, 267:68870, 2025.

Zequn Yang, Yake Wei, Haotian Ni, Zhihao Xu, and Di Hu. Information-theoretic decomposition for multimodal interaction learning. *arXiv preprint arXiv:2606.11614*, 2026.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. Multimodal classification via total correlation maximization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MbQhdzAhSl.

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.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pp. 41753–41769. PMLR, 2023.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Qinghua Hu, Cheng Deng, Cai Xu, Jie Wen, Di Hu, et al. Multimodal fusion on low-quality data: A comprehensive survey. *Information Fusion*, pp. 104437, 2026.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27446–27456. IEEE, 2024.

A Supplementary Experiments

A.1 Datasets

We evaluate our method on seven datasets covering audio–visual classification, image–text classification, and trimodal humor and sentiment recognition. For CREMA-D, AVE, MVSA-Single, and UPMC-Food101, we follow the data preprocessing protocols used in MoNet (Feng et al., 2026). For Kinetics-Sounds, we use the processed data provided by DGL (Wei et al., 2025), while for UR-FUNNY, we follow the preprocessing protocol used in MASAM (Chen et al., 2026). For CMU-MOSEI, we use preprocessed aligned language, acoustic, and visual features for binary sentiment classification, excluding examples with zero-valued sentiment labels from all splits. Table 5 summarizes the dataset characteristics and split information.

Table 5: Dataset characteristics and split information.

Modalities	Dataset	Classes	Train / Val / Test
Audio–Visual	CREMA-D	6	6,697 / 744 / 744
	AVE	28	3,312 / 402 / 401
	Kinetics-Sounds	33	20,852 / 2,319 / 1,633
Image–Text	MVSA-Single	3	1,555 / 519 / 518
	UPMC-Food101	101	62,971 / 22,715 / 5,000
Language–Audio–Visual	UR-FUNNY	2	7,614 / 980 / 994
	CMU-MOSEI	2	16,326 / 1,871 / 4,659

CREMA-D (Cao et al., 2014) is an audio–visual emotion recognition dataset containing utterances performed by 91 actors. Each utterance is associated with one of six emotions: anger, disgust, fear, happiness, sadness, and neutrality. We use speech spectrograms and video frames as the two modalities to predict the emotion of each utterance.

AVE (Tian et al., 2018) contains ten-second videos covering 28 audio–visual event categories, including human activities, animal sounds, and musical performances. Originally introduced for audio–visual event localization, it provides paired audio and visual observations of events in unconstrained videos. In our experiments, we use the dataset for event classification from audio spectrograms and video frames.

Kinetics-Sounds (Arandjelovic & Zisserman, 2017) is a subset of Kinetics containing action categories with both audible and visible cues, such as playing musical instruments. The resulting dataset contains 33 categories, with 20,852 training, 2,319 validation, and 1,633 test clips.

MVSA (Niu et al., 2016) is a multimodal sentiment dataset collected from Twitter, where each example pairs an image with its accompanying text. We use the MVSA-Single variant and perform three-class sentiment classification into positive, neutral, and negative categories. The image and text provide complementary observations of the sentiment expressed in a social media post.

UPMC-Food101 (Wang et al., 2015) is an image–text food recognition dataset with 101 categories. Each example combines a food image with textual information collected from its associated web page. We use both modalities to predict the food category, distinguishing this multimodal benchmark from the image-only Food-101 dataset.

UR-FUNNY (Hasan et al., 2019) is a multimodal humor recognition dataset derived from TED talks. It provides language, acoustic, and visual information for distinguishing humorous from non-humorous utterances, with humor annotations based on audience laughter.

CMU-MOSEI (Zadeh et al., 2018) contains opinion utterances from online videos with language, acoustic, and visual features. Its sentiment annotations range from -3 to $+3$. We use the preprocessed aligned features and formulate the task as binary sentiment classification: examples with negative scores are labeled negative, examples with positive scores are labeled positive, and examples with zero scores are excluded from all splits.

Table 6: Training configurations across the seven datasets. Uni. and Fusion denote unimodal training and fusion training, respectively. β is the cross-fold gain regularization coefficient, and K is the number of folds. Unimodal epochs are reported per fold.

Dataset	Optimizer	Learning rate		β	K	Epochs	
	Uni. / Fusion	Uni.	Fusion			Uni.	Fusion
CREMA-D	SGD / Adam	1e-2	3e-3	0.1	5	140	100
AVE	SGD / Adam	1e-2	3e-3	0.5	5	100	100
Kinetics-Sounds	SGD / Adam	1e-2	3e-3	0.1	5	100	100
MVSA-Single	SGD / Adam	1e-2	3e-3	0.1	5	70	100
UPMC-Food101	SGD / Adam	1e-2	3e-3	0.1	5	70	100
UR-FUNNY	Adam / Adam	1e-4	3e-3	0.1	5	70	100
CMU-MOSEI	Adam / Adam	1e-4	3e-3	0.1	5	50	100

A.2 Training Configurations

Table 6 summarizes the optimization settings for unimodal training and fusion training on each dataset. The unimodal training budget applies to each fold independently.

A.3 Baseline Methods

We compare our method with three standard fusion baselines and nine representative multimodal learning methods, as summarized below.

- **Summation** jointly trains modality encoders and combines their features through element-wise summation for classification.
- **Concatenation** jointly trains modality encoders and concatenates their features as input to a fusion classifier.
- **UniAvg** trains each unimodal model independently and averages their classification logits at inference, without learning an additional fusion module.
- **OGM-GE** (Peng et al., 2022) adaptively modulates modality-specific gradients according to their relative contributions to the learning objective. It additionally injects Gaussian noise to mitigate the potential loss of generalization caused by gradient modulation.
- **AGM** (Li et al., 2023) estimates modality contributions through Shapley-based attribution and adaptively modulates their gradients. Its formulation supports multiple modalities and nonlinear fusion architectures.
- **MMPareto** (Wei & Hu, 2024) integrates gradients from unimodal and multimodal objectives using Pareto optimization. It adjusts both gradient direction and magnitude to reduce optimization conflicts and provide unimodal assistance.
- **MLA** (Zhang et al., 2024) alternates unimodal adaptation while continuously optimizing a shared classification head with gradient modification to preserve previously learned information. It combines modality predictions through uncertainty-based fusion at inference.
- **DGL** (Wei et al., 2025) decouples encoder optimization from fusion optimization by routing unimodal losses to modality encoders and the multimodal loss to the fusion module, reducing interference between their learning objectives.
- **InfoReg** (Huang et al., 2025) regulates information acquisition during the early training stage. It slows information acquisition in information-rich modalities to facilitate learning in information-poor modalities.
- **MASAM** (Chen et al., 2026) adaptively applies sharpness-aware minimization according to modality dominance, estimated from convergence speed and gradient alignment. It uses modality-decoupled perturbation scaling to reduce cross-modal interference.

- **TCNE** (Yu et al., 2026) estimates a lower bound on the total correlation among modality representations and labels. The resulting TCMax objective maximizes this bound to alleviate modality competition and encourage cross-modal alignment.
- **MoNet** (Feng et al., 2026) organizes multimodal training into nested optimization processes. It decouples modality-specific learning into independent optimization streams and coordinates fusion through memories operating at multiple timescales.

B Details of the Preliminary Analysis

This appendix derives the optimization and statistical identities used in Sections 3.2 and 3.3, and specifies the intervention underlying Fig. 2. The directional analysis concerns affine fusion with fixed unimodal encoders and classifiers. It describes a mechanism by which seen-sample fitting can harm generalization; it does not assert that this mechanism occurs for every dataset or every fusion architecture.

B.1 Parameterization and Class Centering

Let $C \geq 2$, $\mathbf{z}^{(m)}(\mathbf{x}) \in \mathbb{R}^{d_m}$, $A_m, W_m \in \mathbb{R}^{C\times d_m}$, and $\mathbf{a}_m, \mathbf{b} \in \mathbb{R}^C$. Features and parameters in the directional analysis are finite. We identify $\theta = (W_1, \ldots, W_M, \mathbf{b})$ with a vector by concatenating its entries. For parameter tuples θ and $D = (B_1, \ldots, B_M, \mathbf{d})$, the inner product and induced norm are

$$\langle \theta, D\rangle = \sum_{m=1}^{M} \operatorname{tr}(W_m^\top B_m) + \mathbf{b}^\top \mathbf{d}, \qquad \|D\|^2 = \sum_{m=1}^{M} \|B_m\|_{\mathrm{F}}^2 + \|\mathbf{d}\|_2^2. \tag{17}$$

All encoders and unimodal classifiers remain fixed throughout this analysis, so each classifier direction D_m is fixed. Softmax probabilities are invariant to adding the same scalar to every class logit. Raw parameter projections can therefore include changes that have no predictive effect. The diagnostic implementation removes this ambiguity by centering every weight matrix and bias across classes. Define

$$\Pi = I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top, \qquad A_m^{\mathrm{c}} = \Pi A_m, \quad \mathbf{a}_m^{\mathrm{c}} = \Pi \mathbf{a}_m, \quad W_m^{\mathrm{c}} = \Pi W_m, \quad \mathbf{b}^{\mathrm{c}} = \Pi \mathbf{b}. \tag{18}$$

For any logits $\mathbf{s}$, $\Pi\mathbf{s} = \mathbf{s} - C^{-1}(\mathbf{1}^\top\mathbf{s})\mathbf{1}$. Hence,

$$\operatorname{softmax}(\Pi\mathbf{s}) = \operatorname{softmax}(\mathbf{s}), \qquad \ell(\Pi\mathbf{s}, y) = \ell(\mathbf{s}, y), \qquad (\Pi\mathbf{s})_y - (\Pi\mathbf{s})_k = s_y - s_k. \tag{19}$$

Centering preserves classification margins, probabilities, and predictions. Unless stated otherwise, the parameter projections below use these centered representatives, with the superscript c suppressed. Their associated logits are likewise centered: $\mathbf{s}^{(m),\mathrm{c}} = \Pi\mathbf{s}^{(m)}$ and $\mathbf{s}_\theta^{\mathrm{c}} = \Pi\mathbf{s}_\theta$. Thus, identities involving logits and centered parameter directions are understood in the same centered representation. This is a convention for measuring saved parameters, not a change to the training algorithm.

Connection to the diagnostic implementation. The diagnostic code uses a fixed average of unimodal logits plus a scaled affine residual:

$$\begin{aligned} \bar{\mathbf{s}}(\mathbf{x}) &= \frac{1}{M}\sum_{m=1}^{M} \big(A_m \mathbf{z}^{(m)}(\mathbf{x}) + \mathbf{a}_m\big), \\ \mathbf{s}_\theta(\mathbf{x}) &= \bar{\mathbf{s}}(\mathbf{x}) + \rho\left(\sum_{m=1}^{M} V_m \mathbf{z}^{(m)}(\mathbf{x}) + \mathbf{d}\right), \end{aligned} \tag{20}$$

where $\rho > 0$ is fixed. This predictor has exactly the affine form in Eq. equation 2, with effective parameters

$$W_m = \frac{A_m}{M} + \rho V_m, \qquad \mathbf{b} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{a}_m + \rho\mathbf{d}. \tag{21}$$

Since the average-logit term is fixed, effective parameter changes satisfy $\theta_t - \theta_0 = \rho(V_{1,t} - V_{1,0}, \ldots, V_{M,t} - V_{M,0}, \mathbf{d}_t - \mathbf{d}_0)$. Applying class centering gives precisely the effective parameter changes measured by the diagnostic code. Thus, the following identities apply both to a directly parameterized affine fusion head and to the affine residual experiment. The reference is the saved training initialization, which need not reproduce logit averaging.

B.2 Directional Incentive of the Fusion Objective

For a nonempty fusion dataset S, let $\mathcal{L}_S(\theta) = |S|^{-1} \sum_{(\mathbf{x},y)\in S} \ell(\mathbf{s}_\theta(\mathbf{x}), y)$, where $\ell(\mathbf{s}, y) = \log \sum_{j=1}^{C} \exp(s_j) - s_y$. The classifier of modality m defines $D_m = (0, \ldots, A_m, \ldots, 0, \mathbf{a}_m)$, with A_m occupying the W_m block. Affinity gives the exact identity

$$\mathbf{s}_{\theta+\alpha D_m}(\mathbf{x}) = \mathbf{s}_\theta(\mathbf{x}) + \alpha \mathbf{s}^{(m)}(\mathbf{x}), \qquad \mathbf{s}^{(m)}(\mathbf{x}) = A_m \mathbf{z}^{(m)}(\mathbf{x}) + \mathbf{a}_m. \tag{22}$$

Proposition B.1 (Directional derivative of fusion cross-entropy)**.** *For fixed unimodal models and finite logits, define $p_{\theta,k}(\mathbf{x}) = [\mathrm{softmax}(\mathbf{s}_\theta(\mathbf{x}))]_k$. Then*

$$\begin{aligned} I_m(\theta; S) &:= -\langle \nabla_\theta \mathcal{L}_S(\theta), D_m \rangle \\ &= \frac{1}{|S|} \sum_{(\mathbf{x},y)\in S} \sum_{k\neq y} p_{\theta,k}(\mathbf{x}) \big(s_y^{(m)}(\mathbf{x}) - s_k^{(m)}(\mathbf{x}) \big). \end{aligned} \tag{23}$$

If modality m strictly separates every sample in S, then $I_m(\theta; S) > 0$ at every finite θ.

Proof. Let $\mathbf{e}_y$ denote the yth coordinate vector. Differentiating cross-entropy with respect to its logits gives

$$\nabla_{\mathbf{s}} \ell(\mathbf{s}, y) = \mathrm{softmax}(\mathbf{s}) - \mathbf{e}_y. \tag{24}$$

Consequently, the derivatives for one sample are

$$\nabla_{W_m} \ell = (\mathbf{p}_\theta - \mathbf{e}_y)(\mathbf{z}^{(m)})^\top, \qquad \nabla_{\mathbf{b}} \ell = \mathbf{p}_\theta - \mathbf{e}_y. \tag{25}$$

Taking the inner product with D_m and averaging yields

$$\begin{aligned} -\langle \nabla_\theta \mathcal{L}_S, D_m \rangle &= \frac{1}{|S|} \sum_{(\mathbf{x},y)\in S} (\mathbf{e}_y - \mathbf{p}_\theta)^\top (A_m \mathbf{z}^{(m)} + \mathbf{a}_m) \\ &= \frac{1}{|S|} \sum_{(\mathbf{x},y)\in S} \left(s_y^{(m)} - \sum_{k=1}^{C} p_{\theta,k} s_k^{(m)} \right) \\ &= \frac{1}{|S|} \sum_{(\mathbf{x},y)\in S} \sum_{k\neq y} p_{\theta,k} (s_y^{(m)} - s_k^{(m)}). \end{aligned} \tag{26}$$

For finite logits, $p_{\theta,k} > 0$. Under strict separation, $s_y^{(m)} - s_k^{(m)} > 0$ for every sample and every $k \neq y$, so the sum is strictly positive. Class centering does not change this expression because $(\mathbf{e}_y - \mathbf{p}_\theta)^\top \mathbf{1} = 0$. □

The strict-separation condition can be relaxed without approximation. Writing $\gamma_{m,k}(\mathbf{x}, y) = s_y^{(m)}(\mathbf{x}) - s_k^{(m)}(\mathbf{x})$, Eq. equation 23 is positive exactly when

$$\sum_{(\mathbf{x},y)\in S} \sum_{\substack{k\neq y \\ \gamma_{m,k}>0}} p_{\theta,k} \gamma_{m,k} > \sum_{(\mathbf{x},y)\in S} \sum_{\substack{k\neq y \\ \gamma_{m,k}<0}} p_{\theta,k} |\gamma_{m,k}|. \tag{27}$$

This is a condition on fusion-weighted unimodal margins. Unimodal accuracy or average cross-entropy alone does not determine its magnitude or establish an ordering between modalities.

The role of training membership. Equation equation 23 holds for either seen or unseen samples; training membership does not change the derivative formula. Its relevance is that unimodal fitting can make margins predominantly positive on seen samples, allowing Eq. equation 27 to hold repeatedly during fusion training. Unseen samples can expose unimodal errors: if an incorrect class k has a larger unimodal logit than the true class, its term in Eq. equation 23 is negative and opposes reinforcement. Such errors do not necessarily make the total alignment negative. Under near-separation, positivity must be checked at each iterate because the fusion probabilities change during training. This distinguishes a possible membership-dependent optimization mechanism from a guarantee that seen and unseen training favor opposite directions.

From a gradient direction to a parameter update. Assume $D_m \neq 0$ and define the cumulative normalized projection $q_m(t) = \langle \theta_t - \theta_0, D_m/\|D_m\| \rangle$. For ordinary gradient descent on $\mathcal{L}_S$, the update $\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \mathcal{L}_S(\theta_t)$ implies

$$q_m(t+1) - q_m(t) = \frac{\eta_t}{\|D_m\|} I_m(\theta_t; S). \tag{28}$$

This equality is exact and does not require a small-step approximation. For a fixed modality and learning rate, a larger I_m gives a larger increment in q_m. Comparing increments across modalities additionally requires the factor $\|D_m\|^{-1}$. Positive alignment describes the component of the update along D_m; the remaining update need not vanish.

For a minibatch update, the same identity holds with that minibatch replacing S. Under strict separation of every training sample, each minibatch has positive alignment. Under only the aggregate condition in Eq. equation 27, individual minibatches need not have positive alignment. Neither case guarantees that the projection grows without bound, since its increments can tend to zero.

When training uses stochastic input augmentation, the minibatch identity applies to the actual transformed inputs in that update. Strict separation of one cached view of each sample does not establish strict separation of all augmented views.

Equation equation 28 concerns unregularized Euclidean gradient descent in the effective parameterization. If gradient descent is performed on $(V_1, \ldots, V_M, \mathbf{d})$ in Eq. equation 20, the effective learning rate is $\eta_t \rho^2$. Momentum, adaptive preconditioning, weight decay, or an additional loss term can change the actual update direction. For these optimizers, the diagnostic measures $\langle \theta_{t+1} - \theta_t, D_m/\|D_m\| \rangle$ directly rather than identifying it with the negative cross-entropy gradient.

B.3 Projection Removal and Unseen-Sample Errors

Let $\Delta\theta_t = \theta_t - \theta_0$ and define

$$\alpha_m(t) = \frac{\langle \Delta\theta_t, D_m \rangle}{\|D_m\|^2} = \frac{q_m(t)}{\|D_m\|}, \qquad \Delta\theta_{t,m}^{\perp} = \Delta\theta_t - \alpha_m(t) D_m. \tag{29}$$

By construction, $\langle \Delta\theta_{t,m}^{\perp}, D_m \rangle = 0$. The intervention removes the accumulated component along D_m while preserving the initialization and the orthogonal remainder:

$$\theta_t^{-m} = \theta_0 + \Delta\theta_{t,m}^{\perp} = \theta_t - \alpha_m(t) D_m, \qquad \mathbf{s}_{\theta_t^{-m}}(\mathbf{x}) = \mathbf{s}_{\theta_t}(\mathbf{x}) - \alpha_m(t) \mathbf{s}^{(m)}(\mathbf{x}). \tag{30}$$

In the residual implementation, the fixed average-logit predictor remains unchanged; the corresponding subtraction from the residual parameters is $\alpha_m(t) D_m/\rho$. No retraining is performed.

For a sample with label y and a competing class $k \neq y$, Eq. equation 30 gives

$$s_{\theta_t, y} - s_{\theta_t, k} = \left(s_{\theta_t^{-m}, y} - s_{\theta_t^{-m}, k} \right) - \alpha_m(t) \left(s_k^{(m)} - s_y^{(m)} \right). \tag{31}$$

Suppose $\alpha_m(t) > 0$ and modality m favors k over y. Then its accumulated component reduces the fused true-class margin against k. If the predictor after removal correctly classifies the sample with a strict margin, adding the component makes the prediction incorrect whenever

$$\alpha_m(t) \left(s_k^{(m)} - s_y^{(m)} \right) > s_{\theta_t^{-m}, y} - s_{\theta_t^{-m}, k} > 0. \tag{32}$$

All remaining learned contributions are already included in the right-hand margin. The condition therefore allows compensation by other modalities, but identifies when that compensation is insufficient. Merely observing $\alpha_m > 0$ or a unimodal error is not enough to conclude that the fused prediction changes.

The removal experiment isolates a parameter component of the fitted predictor. It does not identify the model that would result from training without that component. Moreover, classifier directions need not be mutually orthogonal: for $m \neq n$, $\langle D_m, D_n \rangle = \mathbf{a}_m^\top \mathbf{a}_n$, due to the shared fusion bias. We therefore analyze each removal separately; the recovered-error counts and projection energies are not additive across modalities. The projections are measured in the fixed representation coordinates of the trained encoders and are not invariant to arbitrary feature reparameterizations.

Table 7: Normalized cross-entropy descent incentives at the final seen-trained fusion head, evaluated on fixed diagnostic caches. A positive value favors increasing the corresponding classifier-direction component.

Diagnostic set	Samples	Audio	Video
Seen	670	1.08×10^{-5}	7.22×10^{-7}
Unseen	670	-0.29763	0.17138
Validation	744	-0.27427	0.16284

B.4 Diagnostic Protocol and Numerical Accounting

The CREMA-D diagnostic loads the pretrained unimodal models, freezes them in evaluation mode, and trains new affine fusion heads. The seen and unseen arms use the same saved head initialization and equally sized fusion subsets, each containing 670 samples. The classifier directions are constructed from the pretrained linear heads and checked against their cached unimodal logits. Both directions and fusion parameters are class-centered before computing projections. Parameter changes are measured relative to the exact saved initialization and include the fixed residual scale. Each actual optimizer update is recorded, including any effects of optimizer state and weight decay. Training uses the live training loader, including its input augmentation, whereas checkpoint diagnostics use fixed cached views of the seen, unseen, and validation samples. Predictions before and after removal use identical cached features. The cache-based cross-entropy derivatives below are therefore distinguished from gradients on the augmented training minibatches.

For a specified numerical tolerance $\tau \geq 0$, an update is counted as positive when $q_m(t+1) - q_m(t) > \tau$, negative when it is below $-\tau$, and near-zero otherwise. The tolerance is recorded by the diagnostic protocol. Projection and removal calculations use double precision after the saved features, logits, and parameters are loaded.

The figure discussed in the main text uses the seen arm. Its cumulative projections are $q_a = 2.73159$ and $q_v = 0.52536$ at the final checkpoint, and all 1,100 recorded audio projection increments are positive under the diagnostic tolerance. These are observations about the recorded optimization trajectory, not consequences asserted for an arbitrary optimizer.

Directional incentives at the same fitted head. To connect Eq. equation 23 to the observed trajectory, Table 7 reports $I_m(\theta_t; S)/\|D_m\|$ on each diagnostic cache at the final seen-trained checkpoint. All entries use the same unimodal models and fusion parameters; they do not compare heads trained on different subsets. The audio incentive is positive on seen samples and negative on both unseen and validation samples. Since

$$\frac{d}{d\epsilon}\mathcal{L}_S\left(\theta_t + \epsilon\frac{D_m}{\|D_m\|}\right)\Big|_{\epsilon=0} = -\frac{I_m(\theta_t; S)}{\|D_m\|}, \tag{33}$$

a sufficiently small displacement in the positive audio direction decreases seen-cache cross-entropy while increasing unseen-cache and validation-cache cross-entropy at this checkpoint. This is a local directional observation, not a statement about the complete optimizer update or a population-risk guarantee. The removal experiment below evaluates the finite accumulated component instead.

Let V be the validation set, $n = |V|$, and let $\widehat{y}_{\text{avg}}$, $\widehat{y}$, and $\widehat{y}^{-a}$ denote predictions from averaged unimodal logits, full fusion, and fusion after audio-component removal, respectively. The fusion-induced error set and its recovered subset are

$$\begin{aligned} \mathcal{E}_{\text{fus}} &= \{i : \widehat{y}_{\text{avg},i} = y_i,\ \widehat{y}_i \neq y_i\}, \\ \mathcal{E}_{\text{rec}} &= \{i \in \mathcal{E}_{\text{fus}} : \widehat{y}_i^{-a} = y_i\}. \end{aligned} \tag{34}$$

Thus, $|\mathcal{E}_{\text{rec}}|/|\mathcal{E}_{\text{fus}}|$ measures recovery specifically among errors introduced relative to logit averaging. Overall accuracy change instead counts all repaired and newly introduced errors:

$$\begin{aligned} n\left(\text{Acc}^{-a} - \text{Acc}\right) = &\left|\{i : \widehat{y}_i \neq y_i,\ \widehat{y}_i^{-a} = y_i\}\right| \\ &- \left|\{i : \widehat{y}_i = y_i,\ \widehat{y}_i^{-a} \neq y_i\}\right|. \end{aligned} \tag{35}$$

In particular, $28/43$ and the 4.17-point improvement summarize different quantities. Removal repairs 38 errors in total, of which 28 were introduced by fusion relative to logit averaging, and intro-

Table 8: Final-checkpoint component-removal diagnostic on the 744 CREMA-D validation samples. Counts and accuracies refer to the seen-trained fusion head.

Quantity	Value
Correct predictions from logit averaging	598 / 744
Correct predictions from full fusion	559 / 744
Correct predictions after audio-component removal	590 / 744
Fusion-induced errors relative to logit averaging	43
Recovered fusion-induced errors	28 / 43
All incorrect predictions repaired by removal	38
Correct predictions broken by removal	7
Net accuracy improvement	4.17 percentage points

duces 7 new errors. Its net improvement is therefore $(38-7)/744 = 4.1667\%$, raising accuracy from $559/744 = 75.13\%$ to $590/744 = 79.30\%$. The original average-logit predictor achieves $598/744 = 80.38\%$, so removal does not recover the entire performance gap. These results support a contribution of the audio-direction component to failure on this run, rather than an exhaustive explanation of all fusion errors.

B.5 Risk Estimation from Independent Unseen Samples

Write $Z = (\mathbf{x}, y) \sim P$, where P is the distribution of input–label pairs. Let $\mathscr{T}$ be the sigma-algebra generated by the data and randomness used to fit and select the unimodal models. The fitted parameters $\widehat{\boldsymbol{\phi}}$ are measurable with respect to $\mathscr{T}$. Fix a fusion head θ, or more generally allow it to be measurable with respect to $\mathscr{T}$, without using the unseen evaluation samples. Assume the loss below is jointly measurable in the training outcome and the input–label pair. Define

$$L_{\theta,\widehat{\boldsymbol{\phi}}}(Z) = \ell\Big(\mathbf{s}_\theta(\mathbf{x}; \widehat{\boldsymbol{\phi}}), y\Big), \qquad R_{\widehat{\boldsymbol{\phi}}}(\theta) = \int L_{\theta,\widehat{\boldsymbol{\phi}}}(z)\, P(dz). \tag{36}$$

Conditional on $\mathscr{T}$, this risk evaluates the fixed fitted predictor on an independent input–label pair. The main text suppresses the subscript $\widehat{\boldsymbol{\phi}}$ in R. No distribution over features is substituted for P.

Proposition B.2 (Unbiasedness and consistency for a fixed fitted predictor)**.** *Let $Z_1, Z_2, \ldots$ be i.i.d. from P and independent of $\mathscr{T}$. Suppose $\mathbb{E}_{Z\sim P}[|L_{\theta,\widehat{\boldsymbol{\phi}}}(Z)|] < \infty$ for almost every realization of $\mathscr{T}$. For $U_q = \{Z_1, \ldots, Z_q\}$, define $\widehat{R}_{U_q}(\theta) = q^{-1}\sum_{i=1}^q L_{\theta,\widehat{\boldsymbol{\phi}}}(Z_i)$. Then*

$$\mathbb{E}\Big[\widehat{R}_{U_q}(\theta) \mid \mathscr{T}\Big] = R_{\widehat{\boldsymbol{\phi}}}(\theta), \qquad \widehat{R}_{U_q}(\theta) \xrightarrow[q\to\infty]{\text{a.s.}} R_{\widehat{\boldsymbol{\phi}}}(\theta). \tag{37}$$

The convergence also holds with conditional probability one for almost every realization of $\mathscr{T}$.

Proof. Independence implies $\mathrm{Law}(Z_i \mid \mathscr{T}) = P$. Once $\mathscr{T}$ is fixed, the function $L_{\theta,\widehat{\boldsymbol{\phi}}}$ is fixed. Therefore,

$$\begin{aligned}\mathbb{E}\Big[\widehat{R}_{U_q}(\theta) \mid \mathscr{T}\Big] &= \frac{1}{q}\sum_{i=1}^q \int L_{\theta,\widehat{\boldsymbol{\phi}}}(z)\, P(dz) \\ &= R_{\widehat{\boldsymbol{\phi}}}(\theta).\end{aligned} \tag{38}$$

Conditional on $\mathscr{T}$, the losses are i.i.d. with finite absolute mean. The strong law of large numbers gives almost-sure convergence to their conditional mean, which equals $R_{\widehat{\boldsymbol{\phi}}}(\theta)$. Since this conditional conclusion holds for almost every $\mathscr{T}$, it also holds almost surely under the joint law. □

For deterministic θ, the population risk is a function of $\widehat{\boldsymbol{\phi}}$, so the tower property further gives the main-text identity

$$\mathbb{E}\Big[\widehat{R}_{U_q}(\theta) \mid \widehat{\boldsymbol{\phi}}\Big] = \mathbb{E}\Big[R_{\widehat{\boldsymbol{\phi}}}(\theta) \mid \widehat{\boldsymbol{\phi}}\Big] = R_{\widehat{\boldsymbol{\phi}}}(\theta). \tag{39}$$

If θ also depends on auxiliary training randomness, conditioning on both $\widehat{\boldsymbol{\phi}}$ and θ gives the corresponding statement. The independence requirement includes model selection: U_q must not have been used to select unimodal checkpoints or other components of the fixed predictor.

B.6 Why Seen Samples and Fitted Fusion Heads Require Different Arguments

For a seen sample Z_i used to construct $\widehat{\phi}$, its conditional distribution need not remain P. For deterministic θ,

$$\mathbb{E}\Big[L_{\theta,\widehat{\phi}}(Z_i) \mid \widehat{\phi}\Big] = \int L_{\theta,\widehat{\phi}}(z)\, P_{Z_i|\widehat{\phi}}(dz), \tag{40}$$

which need not equal the integral against P in Eq. equation 36. This difference concerns dependence between a sample and its fitted predictor, even if all original samples were drawn i.i.d. from the same distribution.

A multiclass example establishes that the identity can fail. Let X have a continuous distribution and let Y be independent of X and uniform on $\{1, \dots, C\}$. Given a finite training set T, define an encoder that returns $\mathbf{e}_{Y_i}$ at each observed input X_i and $\mathbf{e}_1$ at every other input. Training inputs are distinct almost surely. This construction can be embedded in the M-modality model by using one active encoder with $d_m = C$ and setting the other fusion weight blocks to zero. Set the active fusion weight block to λI_C and the fusion bias to zero, where $\lambda > 0$ is fixed independently of T. The empirical risk on T and the population risk of this same fitted predictor are

$$\widehat{R}_T = \log(e^\lambda + C - 1) - \lambda, \qquad R_T = \log(e^\lambda + C - 1) - \frac{\lambda}{C}, \qquad R_T - \widehat{R}_T = \lambda\left(1 - \frac{1}{C}\right) > 0. \tag{41}$$

An independent input coincides with a training input with probability zero, which gives the stated population risk. Class centering leaves these losses unchanged. This construction establishes the absence of a general seen-sample unbiasedness guarantee; it is not a model of CREMA-D or an assumption required by Proposition B.1.

Proposition B.2 is pointwise in the fixed fusion head. Substituting a head $\widehat{\theta}(U_q)$ selected or optimized using U_q removes the independence needed to treat the loss function as fixed. The proposition therefore establishes the suitability of independent unseen samples for risk estimation, not a guarantee that empirical fusion optimization improves population risk or accuracy. Almost-sure convergence also does not imply monotone improvement with every increase in q. Throughout this limit, the fitted predictor is held fixed; the proposition does not compare predictors retrained after changing the unimodal–fusion split ratio.

Finally, OOF construction applies sample exclusion separately to each fitted unimodal model. For a value-independent partition of i.i.d. data, each held-out fold is independent of the data used to fit its corresponding models, provided that the fold is also excluded from model selection. However, records from different folds generally depend on overlapping training sets, and a shared fusion head is fitted using the pooled records. The proposition does not treat those records as N independent evaluations of a fixed head. Label-stratified partitions introduce additional sampling structure, so applying the exact i.i.d. identity to such folds requires accounting for that structure. We use the independent-holdout result to motivate the OOF sample-exclusion principle, without asserting an i.i.d. pooled-OOF bound or guaranteed generalization gains.