# Strong-Modality Collapse in Multimodal Fusion: Causal Analysis and Structural Solution

**Anonymous authors**
Paper under double-blind review

## Abstract

Multimodal fusion is expected to benefit from combining complementary signals. However, causal pathway isolation shows that standard symmetric cross-attention can degrade the predictive performance of the dominant modality, in some cases pushing the fusion model below its unimodal baseline. We call this failure *strong-modality collapse*. We find that it results from the interaction of two factors: cross-modal attention alters the dominant representation, while joint optimization adapts its classifier to these changes. In most configurations, neither factor alone causes substantial degradation, but together they produce losses of up to 42.6 percentage points. This finding motivates Inverted Asymmetric Fusion (IAF), which shields the dominant modality from cross-modal attention while allowing weaker modalities to attend to it. Because structural shielding alone cannot address initial feature imbalances, we introduce Modality-Aware Knowledge Distillation (MAKD) to strengthen weaker encoders before fusion. Across four multimodal benchmarks, IAF preserves the dominant pathway at its unimodal ceiling and improves fusion performance over the strongest unimodal baseline. [1]

## 1 Introduction

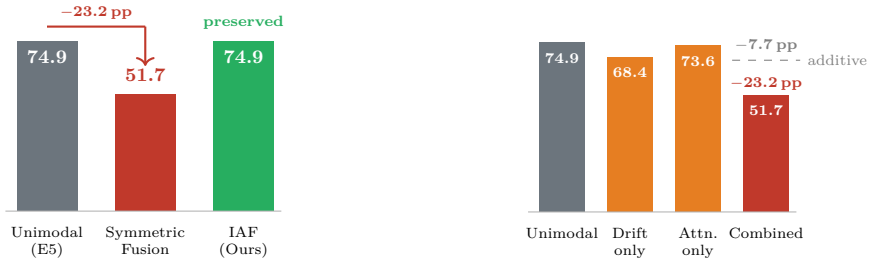
Multimodal fusion is expected to improve performance by combining complementary signals from different modalities (Baltrusaitis et al., 2019; Li & Tang, 2025). Yet models trained for multimodal fusion do not always outperform a model trained on the strongest modality alone. For example, on humor-style classification of videos (Kenneth et al., 2025), an early-fusion model using text, audio, and video is 10.55 percentage points (pp) less accurate than a model using text alone (Table 2). This issue can arise across different fusion strategies, including early fusion, late fusion, and cross-attention-based architectures (Wang et al., 2020b; Huang et al., 2022). Prior work has framed this problem as modality collapse, where a fusion model relies disproportionately on a subset of modalities (Chaudhuri et al., 2025; Javaloy et al., 2022). These studies primarily consider how weaker or noisier modalities fail to contribute to the fused representation. But fusion can fail in another way: it can degrade the modality that is already most predictive of the task. We call this failure *strong-modality collapse*. Using pathway isolation, we evaluate each modality's representation independently of the final fusion output. On the MultiHuSE dataset, for example, the dominant text pathway in a symmetric attention model drops from 74.9% accuracy in the unimodal setting to 51.7% after fusion (Figure 1a).

We next ask why this degradation occurs. Existing work on modality collapse and imbalance has primarily focused on optimization dynamics, showing that modalities overfit and generalize at different rates, such that training favors the fastest-converging modality (Huang et al., 2022; Javaloy et al., 2022; Fan et al., 2023). However, these explanations do not account for cases in which fusion degrades the dominant modality itself. We therefore investigate what happens to the dominant modality when it participates in symmetric

[1] Code: `https://anonymous.4open.science/r/strong_modality_collaspe-3F67`



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.




(a) Symmetric fusion degrades the dominant modality by 23.2 pp; IAF preserves it.

(b) The two causes interact: the combined drop (−23.2) is three times the additive drop (−7.7).

Figure 1: (a) Pathway accuracy before and after Symmetric Fusion, vs. IAF (Table 3). (b) The $2 \times 2$ causal decomposition (Table 4).

cross-attention. Through causal pathway isolation, we decompose the degradation into two factors: cross-modal attention alters the dominant modality's features, while joint optimization allows its classifier to drift to accommodate that change. Neither factor alone causes substantial degradation, but together their effect is much larger than the additive prediction (Figure 1b).

This finding points to a structural solution. If the degradation requires both cross-modal attention and classifier adaptation, removing either should prevent the interaction without requiring any rebalancing of gradients or losses. We propose *Inverted Asymmetric Fusion* (IAF), in which the dominant modality does not attend to other modalities, while weaker modalities still attend to it. The dominant pathway therefore retains its unimodal representation, and its classifier is kept frozen. Since shielding the dominant modality does not strengthen the weaker encoders, we also introduce *Modality-Aware Knowledge Distillation* (MAKD) to strengthen them before fusion. Across four multimodal benchmarks, IAF preserves the dominant pathway at its unimodal ceiling and consistently improves fused accuracy over the strongest unimodal baseline.

To summarize, our contributions are: (i) we identify *strong-modality collapse* and show through causal pathway isolation that it arises from an interaction between cross-modal attention and classifier adaptation; (ii) we propose IAF to shield the dominant modality and MAKD to strengthen weaker modalities before fusion; and (iii) we evaluate the resulting approach across four multimodal benchmarks with different dominant modalities.

## 2 Related Work

**Multimodal Fusion Architectures.** Multimodal fusion has progressed from simple concatenation and prediction averaging to tensor-based interactions (Baltrusaitis et al., 2019; Pu Liang et al., 2021; Zadeh et al., 2017), efficient low-rank factorization (Liu et al., 2018), and cross-modal attention (Tsai et al., 2020). Subsequent work has incorporated pretrained language models (Rahman et al., 2020), bottleneck constraints (Nagrani et al., 2021), and modality subspace disentanglement (Hazarika et al., 2020). More recent approaches explore asymmetric or guided fusion, including text-guided integration (Zhang et al., 2025) and knowledge-guided cross-modal interactions (Feng et al., 2024). However, these approaches do not consider degradation of the dominant modality under cross-modal attention. IAF focuses on addressing this failure.

**Modality Collapse and Pathway Analysis.** Multimodal fusion can fail to use all available modalities, a phenomenon known as modality collapse. Prior work attributes this to differences in overfitting and generalization (Wang et al., 2020b; Wu et al., 2022), conflicting gradients and other optimization dynamics (Javaloy et al., 2022; Huang et al., 2022). Chaudhuri et al. (2025) provide a mechanistic account in which noisy and predictive features become entangled in shared fusion-head neurons under a low-rank gradient bottleneck, and show that cross-modal knowledge distillation can reduce this entanglement.

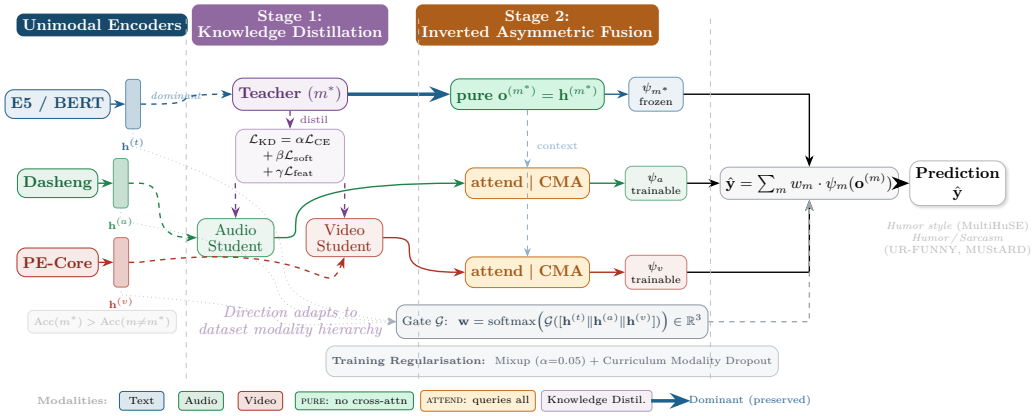



Figure 2: Overview of the proposed two-stage framework. Stage 1 performs Modality-Aware Knowledge Distillation, and Stage 2 applies the Inverted Asymmetric Fusion.

Other approaches address modality imbalance through gradient modulation, Pareto-based optimization, or prototype-guided objectives (Wei et al., 2025; Wei & Hu, 2024; Fan et al., 2023). These studies primarily assess modality use from the fused representation or final output. We instead isolate each modality pathway and evaluate its predictive performance after fusion (Section 3.2), revealing degradation that the final fusion output can hide.

**Knowledge Distillation in Multimodal Learning.** Knowledge distillation transfers information from a teacher to a student by matching predictions or representations (Hinton et al., 2015; Gou et al., 2021) and has been applied to cross-modal tasks including depth estimation, optical flow, action recognition, and emotion recognition (Gupta et al., 2016; Thoker & Gall, 2019; Albanie et al., 2018). In multimodal learning, distillation has been used for model compression, missing-modality robustness, and representational alignment (Fang et al., 2021; Wang et al., 2023; 2020a; Wei et al., 2023; Li et al., 2023b; Lin & Hu, 2024). Our MAKD differs in its placement and role: it is applied before fusion and uses the empirically dominant modality as a teacher to strengthen weaker encoders, complementing IAF, which protects the dominant pathway during fusion. A detailed comparison with existing multimodal distillation methods is provided in Appendix M.

## 3 Diagnosing and Correcting Strong-Modality Collapse

We formalize the multimodal fusion setting and diagnose strong-modality collapse through pathway isolation, causal decomposition, and representational analysis. These analyses motivate IAF, which protects the dominant modality, and MAKD, which strengthens weaker modalities before fusion. Figure 2 shows the resulting pipeline. Although training proceeds sequentially, with MAKD followed by IAF, we present the method in diagnostic order to connect the observed failure to the proposed design.

### 3.1 Problem Formulation

Let $\mathcal{M}$ denote the set of modalities available for a given sample. For each modality $m \in \mathcal{M}$, let $\mathbf{f}^{(m)} \in \mathbb{R}^{d_m}$ denote its pre-extracted feature vector, projected to a shared $d$=512 dimensional space through a modality-specific projection, $\mathbf{h}^{(m)} = \phi_m(\mathbf{f}^{(m)}) \in \mathbb{R}^{512}$. We define the dominant modality set $\mathcal{M}^*$ as the modality or modalities achieving the highest standalone validation accuracy:

$$\mathcal{M}^* = \arg\max_{m \in \mathcal{M}} \mathrm{Acc}_{\mathrm{uni}}(m). \tag{1}$$

### 3.2 Diagnosing Strong-Modality Collapse

Standard fusion evaluation considers only the final prediction, which can hide degradation within individual pathways. We therefore separate pathway behavior from the fused prediction using three analyses: *pathway isolation* tests whether the dominant pathway degrades; *causal decomposition* separates the effects of cross-modal attention and classifier adaptation; and *representational similarity analysis* examines changes in the pathway representation.

**Cross-Modal Anchor.** The *Cross-Modal Anchor* (CMA) is our basic cross-modal attention operation. For a modality $m$, it takes the query vector $\mathbf{h}^{(m)} \in \mathbb{R}^{512}$ and a context matrix $\mathbf{C} = [\mathbf{h}^{(j)}]_{j \neq m} \in \mathbb{R}^{K \times 512}$ that stacks the vectors of the other $K$ modalities. CMA applies standard multi-head attention (MHA) (Vaswani et al., 2017) ($d_{\text{model}}$=512, $H$=8 heads) with $\mathbf{h}^{(m)}$ as the query and $\mathbf{C}$ as keys and values, followed by dropout (rate 0.2), a residual connection, and layer normalization:

$$\text{CMA}(\mathbf{h}^{(m)}, \mathbf{C}) = \text{LayerNorm}\Big(\mathbf{h}^{(m)} + \text{Dropout}\Big(\text{MHA}(\mathbf{h}^{(m)}, \mathbf{C}, \mathbf{C})\Big)\Big). \tag{2}$$

**Symmetric Fusion Setup.** We first consider a conventional symmetric fusion architecture in which each modality attends to all other modalities through CMA. For each modality $m \in \mathcal{M}$, the fused representation is given by

$$\mathbf{o}_{\text{sym}}^{(m)} = \text{CMA}\Big(\mathbf{h}^{(m)}, [\mathbf{h}^{(j)}]_{j \neq m}\Big). \tag{3}$$

This applies cross-modal attention to every modality, including the dominant one, and serves as our reference architecture.

**Pathway Isolation.** To measure whether fusion alters an individual modality's predictive performance, we evaluate each pathway independently of the final aggregation mechanism. For a fused model, let $\mathbf{o}^{(m)}$ denote the post-fusion representation of modality $m$ and $\psi_m$ its corresponding classifier, so that pathway accuracy is $\text{Acc}_{\text{path}}(m) = \text{Acc}(\psi_m(\mathbf{o}^{(m)}), y)$.

For the dominant modality, we compare $\text{Acc}_{\text{path}}(m^*)$ with its standalone accuracy $\text{Acc}_{\text{uni}}(m^*)$. A decrease indicates the dominant pathway has lost predictive performance during fusion, even if the final multimodal prediction remains competitive.

**Causal Decomposition.** Pathway isolation shows that symmetric fusion can degrade the dominant modality but does not identify why. In symmetric cross-attention, two factors change together: cross-modal attention modifies the dominant representation, while joint optimization allows its classifier to adapt. We separate these factors using a $2 \times 2$ design crossing cross-modal attention (*off/on*) with classifier state (*frozen/trainable*). The classifier is initialized identically across conditions. Cell 1 disables attention and freezes the classifier, providing the unimodal reference. Cell 2 enables classifier adaptation only, Cell 3 enables cross-modal attention only, and Cell 4 enables both, corresponding to standard symmetric fusion.

Relative to Cell 1, we define $\Delta_{\text{attn}} = \text{Acc}_1 - \text{Acc}_3$, $\Delta_{\text{drift}} = \text{Acc}_1 - \text{Acc}_2$, and $\Delta_{\text{joint}} = \text{Acc}_1 - \text{Acc}_4$. If the two factors acted independently, $\Delta_{\text{joint}}$ would be approximately additive ($\Delta_{\text{joint}} \approx \Delta_{\text{attn}} + \Delta_{\text{drift}}$). We therefore quantify their interaction as

$$\Delta_{\text{int}} = \Delta_{\text{joint}} - \Delta_{\text{attn}} - \Delta_{\text{drift}}. \tag{4}$$

A positive $\Delta_{\text{int}}$ indicates that the two factors compound rather than contribute independently to the observed collapse.

**Representational Analysis via CKA.** Pathway accuracy measures changes in predictive performance, but not representational change. We therefore complement it with linear Centered Kernel Alignment (CKA) (Kornblith et al., 2019). For each modality, we compare the fused pathway representations with the corresponding standalone unimodal representations on the same evaluation samples, separately for Symmetric Fusion and IAF. CKA values closer to one indicate greater representational similarity, while lower values indicate greater representational drift.

### 3.3 Correcting Strong-Modality Collapse with IAF

To prevent strong-modality collapse, IAF protects the dominant modality from cross-modal attention and freezes its classifier, while weaker modalities attend to it and adapt their classifiers to the resulting context-enriched representations.

**Pure and Attend Pathways.** Let $m^*$ denote the dominant modality (or modalities). IAF imposes an asymmetric constraint on the forward pass:

$$\mathbf{o}^{(m)} = \begin{cases} \mathbf{h}^{(m)} & \text{if } m = m^* \text{ (PURE)}, \\ \mathrm{CMA}\big(\mathbf{h}^{(m)}, [\mathbf{h}^{(j)}]_{j \neq m}\big) & \text{if } m \neq m^* \text{ (ATTEND)}. \end{cases} \tag{5}$$

The PURE pathway bypasses cross-modal attention and retains the dominant modality's original representation. The ATTEND pathways query the other modalities, including the dominant modality, as contextual information. Ablations removing one modality at a time confirm that all modalities contribute to full trimodal performance (Appendix H).

**Adaptive Gating and Final Classification.** Each modality produces pathway logits through a dedicated classifier head $\psi_m$: a frozen copy of the standalone unimodal classifier for the PURE modality, and a trainable linear classifier over the CMA output $\mathbf{o}^{(m)}$ (Eq. 5) for each ATTEND modality. A gating network $\mathcal{G}$ produces sample-adaptive modality weights from the pre-attention representations:

$$\mathbf{w} = \mathrm{softmax}\Big(\mathcal{G}\Big(\big\|_{m \in \mathcal{M}} \mathbf{h}^{(m)}\Big)\Big), \tag{6}$$

where $\|$ denotes concatenation, $\mathbf{w} \in \mathbb{R}^{|\mathcal{M}|}$, and $\mathcal{G}$ is a two-layer MLP with batch normalization and dropout. The final prediction is the weighted sum of pathway logits, $\hat{\mathbf{y}} = \sum_{m \in \mathcal{M}} w_m \, \psi_m(\mathbf{o}^{(m)})$.

Because the gate operates on pre-attention representations, its input does not depend on the modified ATTEND representations, preserving the PURE/ATTEND distinction while letting the final prediction adapt its reliance on each modality per sample. We ablate the gating strategy and analyze learned gate weights in Appendix I and Appendix J.

### 3.4 Training Regularization

We apply two complementary regularization strategies during fusion training: **Mixup**, which interpolates paired feature vectors and their labels within each mini-batch to discourage overconfident predictions (Zhang et al., 2018), and **Curriculum Modality Dropout**, a two-phase stochastic masking schedule that trains all modalities jointly during warm-up before progressively masking modality subsets to prevent over-reliance on any one modality (Morerio et al., 2017). Full formulations and the ablation are in Appendix K.

### 3.5 Strengthening Weak Modalities with MAKD

IAF protects the dominant pathway but does not improve weaker modality representations. We therefore apply MAKD before fusion, using the dominant modality as a teacher to strengthen weaker representations before cross-modal integration.

**Distillation Objective.** Each student $\mathcal{S}$ is trained against a frozen teacher $\mathcal{T}$ with logits $\mathbf{z}^T \in \mathbb{R}^C$ and bottleneck features $\mathbf{g}^T \in \mathbb{R}^{512}$. The objective combines hard-label supervision, soft-label matching, and feature alignment:

$$\mathcal{L}_{\mathrm{KD}} = \alpha \, \mathcal{L}_{\mathrm{CE}} + \beta \, \mathcal{L}_{\mathrm{soft}} + \gamma \, \mathcal{L}_{\mathrm{feat}} \tag{7}$$

where $\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} \mathbf{1}[y = c] \log \mathrm{softmax}(\mathbf{z}^S)_c$ is standard cross-entropy, $\mathcal{L}_{\mathrm{soft}} = T^2 \, \mathrm{KL}\left(\sigma(\mathbf{z}^S/T) \,\|\, \sigma(\mathbf{z}^T/T)\right)$ matches softened output distributions at temperature $T$, and $\mathcal{L}_{\mathrm{feat}} = \|\mathbf{g}^S - \mathbf{g}^T\|_2^2$ aligns bottleneck features. We use $T = 3.5$, $\alpha = 0.4$, $\beta = 0.35$, and $\gamma = 0.25$ across datasets and encoder combinations. A sensitivity analysis of these loss coefficients is provided in Appendix L.

**Dataset-specific teacher configurations.** When multiple modalities are comparably dominant, as on MUStARD, we use weighted dual-teacher distillation. For primary teacher $\mathcal{T}_1$ and secondary teacher $\mathcal{T}_2$,

$$\mathcal{L}_{\text{dual}} = w_1 \mathcal{L}_{\text{KD}}(\mathcal{T}_1) + (1 - w_1)\mathcal{L}_{\text{KD}}(\mathcal{T}_2), \tag{8}$$

where $w_1 = 0.8$ for the audio-primary configuration and $w_1 = 0.2$ for the video-primary configuration.

For UR-FUNNY, the narrow unimodal performance gap motivates same-modality, cross-architecture distillation. High-capacity teachers (E5, Dasheng, and PE-Core) supervise BiLSTM students operating on the benchmark's original GloVe, COVAREP, and OpenFace features. Teacher features are projected to 512 dimensions before feature alignment.

# 4 Experiments

**Datasets.** We evaluate on four multimodal datasets: UPMC Food-101 (Wang et al., 2015) contains 90,688 image-text samples for 101-class food classification, using the standard 67,972/22,716 train-test split from the original paper. MultiHuSE (Kenneth et al., 2025) contains 2,407 video utterances for 5-class humor-style classification, evaluated using 5-fold cross-validation grouped by unique text ID. UR-FUNNY (Kamrul Hasan et al., 2019) contains 16,514 context-punchline pairs for binary humor detection and uses its standard train, validation, and test split. MUStARD (Castro et al., 2019) contains 690 balanced utterances for binary sarcasm detection and uses 5-fold cross-validation.

**Feature Extraction and Evaluation.** All features are extracted offline using frozen pretrained encoders. This keeps the input representations fixed across fusion architectures, so performance differences reflect the fusion mechanism. MultiHuSE, Food-101, and MUStARD, use BERT (Devlin et al., 2019) and Multilingual-E5-large-instruct (Wang et al., 2024) for text, Dasheng (Dinkel et al., 2024) for audio, and VideoMAE (Tong et al., 2022) and PE-Core (Bolya et al., 2025) for video/image. UR-FUNNY uses its original GloVe (Pennington et al., 2014), COVAREP (Degottex et al., 2014), and OpenFace (Baltrusaitis et al., 2018) features for consistency with prior work. We report accuracy and macro-F1, with 5-fold cross-validation for MultiHuSE and MUStARD and five seeds for UR-FUNNY and Food-101. Appendix O provides full training configurations and model parameters.

## 4.1 Establishing Strong-Modality Collapse

**Unimodal Baselines.** Table 1 reports the strongest and second-strongest modality per dataset (raw pretrained features), establishing the dominance hierarchy used as the reference ceiling throughout. Text is dominant on MultiHuSE, Food-101 (except with BERT, where image is dominant), and UR-FUNNY, with a wide margin on MultiHuSE (14.41pp) but a narrow one on Food-101 (0.78pp) and UR-FUNNY (3.7pp). Audio and video are co-dominant on MUStARD, separated by only 1.02pp. MAKD improves weaker unimodal accuracy by 0.54–6.23 pp, except on Food-101, where accuracy decreases by 0.08 and 0.78 pp. Full base-vs-distilled results for all sixteen modality/encoder combinations, including F1, are provided in Appendix C.

Table 1: Mean accuracy of the dominant and runner-up modalities per dataset using raw pretrained features. The dominant accuracy defines the unimodal ceiling. Full results are in Appendix C.

| Dataset | Dominant | Acc (%) | Runner-up | Acc (%) |
|---|---|---|---|---|
| MultiHuSE | Text (E5) | 74.86 | Audio (Dasheng) | 60.45 |
| Food-101 | Text (E5) | 86.24 | Image (PE-Core) | 85.46 |
| UR-FUNNY | Text (GloVe) | 62.11 | Video (OpenFace) | 58.41 |
| MUStARD | Audio (Dasheng) | 78.12 | Video (PE-Core) | 77.10 |

**Standard Fusion.** We next compare Early, Late, and Symmetric Fusion against the unimodal ceiling. Table 2 shows that no fusion strategy consistently improves upon the unimodal ceiling. On MultiHuSE, three of four configurations remain below the ceiling, including a 10.55 pp drop for Early Fusion with E5+VideoMAE. On MUStARD, Symmetric Fusion improves over the audio ceiling in all configurations, while the gains from Early and Late Fusion depend on the encoder pair. On Food-101, all fusion strategies improve over the unimodal baselines, and UR-FUNNY also benefits from all three strategies. These results raise the question of why fusion does not consistently improve over the unimodal ceiling. We therefore examine Symmetric Fusion through pathway isolation to determine how fusion affects individual modality pathways.

Table 2: Baseline fusion accuracy compared with the dominant unimodal ceiling. MultiHuSE and MUStARD report mean ± SD over 5-fold CV; UR-FUNNY and Food-101 report mean ± SD over 5 seeds.

| Dataset | Configuration | Unimodal ceiling | Early | Late | Symmetric |
|---|---|---|---|---|---|
| MultiHuSE | BERT+VideoMAE | 68.26±2.25 (BERT) | 67.22±2.00 | 68.05±2.12 | 69.55±1.74 |
| MultiHuSE | BERT+PE-Core | 68.26±2.25 (BERT) | 70.38±2.57 | 68.63±2.36 | 70.46±2.46 |
| MultiHuSE | E5+VideoMAE | 74.86±3.08 (E5) | 64.31±1.51 | 71.25±4.07 | 72.41±4.81 |
| MultiHuSE | E5+PE-Core | 74.86±3.08 (E5) | 76.28±2.91 | 71.54±4.20 | 74.41±4.36 |
| UR-FUNNY | GloVe+COV+OF | 62.11±1.22 (GloVe) | 64.81±0.77 | 63.86±0.36 | 63.94±0.44 |
| MUStARD | BERT+VideoMAE | 78.12±5.45 (Audio) | 75.94±2.17 | 77.68±3.56 | 79.28±4.26 |
| MUStARD | BERT+PE-Core | 78.12±5.45 (Audio) | 75.94±2.73 | 80.72±4.69 | 82.17±3.57 |
| MUStARD | E5+VideoMAE | 78.12±5.45 (Audio) | 79.57±3.41 | 77.54±4.49 | 79.86±3.09 |
| MUStARD | E5+PE-Core | 78.12±5.45 (Audio) | 79.71±3.04 | 80.72±4.66 | 82.03±2.69 |
| Food-101 | E5+PE-Core | 86.24±0.11 (E5) | 95.42±0.06 | 95.24±0.07 | 94.01±0.18 |
| Food-101 | BERT+PE-Core | 85.46±0.06 (PE-Core) | 92.58±0.11 | 92.74±0.23 | 90.68±0.20 |

### 4.2 Pathway Isolation: Where the Degradation Occurs

Table 3 shows that the dominant pathway falls below its unimodal ceiling in all 11 configurations. The largest drops occur on MultiHuSE BERT+VideoMAE, where the text pathway falls from 68.26% to 32.03%, and Food-101 BERT+PE-Core, where the image pathway falls from 85.46% to 42.89%. In contrast, non-dominant pathways can improve under fusion, for example the MultiHuSE video pathways consistently exceed their unimodal ceilings. These results show that Symmetric Fusion can improve weaker modalities while degrading the strongest modality.

Table 3: Symmetric Fusion pathway accuracy relative to each modality's unimodal ceiling. **Bold** denotes the dominant modality. Values are mean ± SD over 5-fold CV for MultiHuSE and MUStARD, and 5 random seeds for UR-FUNNY and Food-101.

| Dataset | Configuration | Text | | Audio | | Video/Image | |
|---|---|---|---|---|---|---|---|
| | | Ceiling | Symmetric | Ceiling | Symmetric | Ceiling | Symmetric |
| MultiHuSE | BERT+VideoMAE | **68.26±2.25** | **32.03±13.34** | 60.45±1.58 | 57.91±4.47 | 40.13±0.73 | 54.33±14.90 |
| MultiHuSE | BERT+PE-Core | **68.26±2.25** | **53.26±2.93** | 60.45±1.58 | 60.57±5.99 | 43.66±1.29 | 62.86±2.58 |
| MultiHuSE | E5+VideoMAE | **74.86±3.08** | **51.68±9.78** | 60.45±1.58 | 50.97±13.31 | 40.13±0.73 | 66.97±6.48 |
| MultiHuSE | E5+PE-Core | **74.86±3.08** | **53.51±4.03** | 60.45±1.58 | 61.61±7.55 | 43.66±1.29 | 66.47±6.74 |
| UR-FUNNY | GloVe+COV+OF | **62.11±1.22** | **57.83±2.82** | 57.55±2.13 | 53.74±3.23 | 58.41±0.36 | 54.83±3.57 |
| MUStARD | BERT+VideoMAE | 69.42±2.80 | 62.03±4.95 | **78.12±5.45** | **64.35±15.21** | 74.49±1.16 | 67.83±8.02 |
| MUStARD | BERT+PE-Core | 69.42±2.80 | 64.93±9.53 | **78.12±5.45** | **69.42±6.31** | 77.10±3.13 | 73.48±9.01 |
| MUStARD | E5+VideoMAE | 70.29±2.00 | 60.29±12.28 | **78.12±5.45** | **77.25±5.07** | 74.49±1.16 | 71.45±6.18 |
| MUStARD | E5+PE-Core | 70.29±2.00 | 66.96±5.62 | **78.12±5.45** | **71.59±9.27** | 77.10±3.13 | 75.07±3.85 |

| | | Text | | Image | |
|---|---|---|---|---|---|
| | | Ceiling | Symmetric | Ceiling | Symmetric |
| Food-101 | E5+PE-Core | **86.24±0.11** | **78.98±9.33** | 85.46±0.06 | 74.31±8.01 |
| Food-101 | BERT+PE-Core | 76.08±0.03 | 85.38±2.52 | **85.46±0.06** | **42.89±9.39** |

### 4.3 Diagnosing the Collapse

**Causal Decomposition .** Pathway isolation identifies where degradation occurs within symmetric fusion. We next test whether the degradation results from the interaction between

cross-modal attention and classifier adaptation. Table 4 reports the four cells of the $2 \times 2$ design (Section 3.2) for each dominant pathway across all four datasets.

Table 4: Causal decomposition of strong-modality collapse. Dominant-pathway accuracy (%) under a 2×2 grid crossing cross-modal attention (off/on) with classifier state (frozen/trainable). Cell 1 is unimodal reference; Cell 4 is standard Symmetric Fusion.

| Dataset | Config. | Pathway | Cell 1 (no attn, frozen) | Cell 2 (no attn, trainable) | Cell 3 (attn, frozen) | Cell 4 (attn, trainable) |
|---|---|---|---|---|---|---|
| MultiHuSE | bert+videomae | Text | 68.26±2.25 | 66.89±2.81 | 66.89±1.85 | **32.03±13.34** |
| MultiHuSE | bert+PE-Core | Text | 68.26±2.25 | 66.60±2.33 | 67.14±2.47 | **53.26±2.93** |
| MultiHuSE | e5+videomae | Text | 74.86±3.08 | 68.38±15.17 | 73.62±3.32 | **51.68±9.78** |
| MultiHuSE | e5+PE-Core | Text | 74.86±3.08 | 73.82±3.15 | 73.66±3.72 | **53.51±4.03** |
| Food-101 | E5+PE-Core | Text | 86.24±0.11 | 86.45±0.03 | 86.67±0.68 | **78.98±9.33** |
| Food-101 | BERT+PE-Core | Image | 85.46±0.06 | 85.40±0.07 | 85.63±0.16 | **42.89±9.39** |
| MUStARD | bert+videomae | Audio | 78.12±5.45 | 76.96±5.58 | 77.83±5.88 | **64.35±15.21** |
| MUStARD | bert+videomae | Video | 74.49±1.16 | 72.46±2.56 | 73.33±3.05 | **67.83±8.02** |
| MUStARD | bert+PE-Core | Audio | 78.12±5.45 | 77.10±5.90 | 76.67±3.60 | **69.42±6.31** |
| MUStARD | bert+PE-Core | Video | 77.10±3.13 | 72.61±10.35 | 77.10±3.34 | **73.48±9.01** |
| MUStARD | e5+videomae | Audio | 78.12±5.45 | 77.10±5.79 | 77.25±5.77 | **77.25±5.07** |
| MUStARD | e5+videomae | Video | 74.49±1.16 | 72.46±2.23 | 74.20±3.06 | **71.45±6.18** |
| MUStARD | e5+PE-Core | Audio | 78.12±5.45 | 76.52±6.92 | 77.68±5.58 | **71.59±9.27** |
| MUStARD | e5+PE-Core | Video | 77.10±3.13 | 71.59±13.71 | 78.12±2.87 | **75.07±3.85** |
| UR-FUNNY | GloVe+COV+OF | Text | 62.11±1.22 | 48.31±2.77 | 62.98±0.41 | 57.83±2.82 |

Across MultiHuSE, Food-101, and most MUStARD configurations, neither factor alone causes substantial degradation, whereas their combination produces a much larger drop. This interaction is particularly clear on MultiHuSE (Cell 4, up to −36.23 pp) and Food-101 (Cell 4, −42.57 pp). UR-FUNNY differs from this pattern: classifier drift alone causes the largest drop (−13.80 pp), while adding cross-modal attention recovers most of this loss but does not fully restore ceiling performance, leaving a 4.28 pp gap.

**Representational Similarity via CKA.** We further examine changes in the dominant pathway's representation during fusion using linear Centered Kernel Alignment (CKA), comparing each pathway's representation after Symmetric Fusion with its standalone counterpart (1.0 indicates identical representations up to rotation and scaling; full results are provided in Appendix D, Table 9). The dominant text pathway on Food-101 has a CKA of 0.747 in the E5 configuration, while the dominant image pathway has a CKA of 0.962 in the BERT configuration, despite substantial accuracy degradation in both cases. Thus, fusion can alter the dominant representation, but the magnitude of representational change alone does not explain the observed accuracy loss.

### 4.4 Correcting Strong-Modality Collapse

**IAF without MAKD.** We first evaluate IAF on the same raw features as the baseline, without MAKD. As Table 5 shows, IAF exceeds the unimodal ceiling in all eleven configurations, with gains ranging from 0.25 to 8.48 pp. It therefore avoids the below-ceiling performance observed with Symmetric Fusion.

Table 5: Accuracy and improvement over the unimodal ceiling for IAF without MAKD, Symmetric Fusion with MAKD, and IAF with MAKD. Δ denotes fusion accuracy minus the unimodal ceiling. Best-performing results are **bolded.**

| Dataset | Configuration | Unimodal ceiling | IAF | Δ | Symmetric+MAKD | Δ | IAF+MAKD | Δ |
|---|---|---|---|---|---|---|---|---|
| MultiHuSE | BERT+VideoMAE | 68.26±2.25 | 72.29±1.84 | +4.03 | 69.42±1.48 | +1.16 | **73.16±2.59** | +4.90 |
| MultiHuSE | BERT+PE-Core | 68.26±2.25 | 72.33±2.02 | +4.07 | 72.00±1.20 | +3.74 | **74.74±1.71** | +6.48 |
| MultiHuSE | E5+VideoMAE | 74.86±3.08 | 75.11±6.17 | +0.25 | 72.37±2.73 | −2.49 | **76.92±3.53** | +2.06 |
| MultiHuSE | E5+PE-Core | 74.86±3.08 | 75.90±6.17 | +1.04 | 74.32±3.03 | −0.54 | **78.06±3.82** | +3.20 |
| MUStARD | BERT+VideoMAE | 78.12±5.45 | **80.43±2.63** | +2.31 | 78.99±2.59 | +0.87 | 80.29±3.16 | +2.17 |
| MUStARD | BERT+PE-Core | 78.12±5.45 | 80.14±4.39 | +2.02 | 81.59±4.17 | +3.47 | **82.61±3.92** | +4.49 |
| MUStARD | E5+VideoMAE | 78.12±5.45 | 80.14±4.06 | +2.02 | 80.87±2.70 | +2.75 | **81.16±4.20** | +3.04 |
| MUStARD | E5+PE-Core | 78.12±5.45 | 81.74±4.24 | +3.62 | 82.46±4.52 | +4.34 | **83.33±4.20** | +5.21 |
| UR-FUNNY | GloVe+COV+OF | 62.11±1.22 | 63.56±0.29 | +1.45 | 69.18±0.33 | +7.07 | **69.80±0.39** | +7.69 |
| Food-101 | E5+PE-Core | 86.24±0.11 | 94.72±0.18 | +8.48 | **95.29±0.15** | +9.05 | 94.27±0.35 | +8.03 |
| Food-101 | BERT+PE-Core | 85.46±0.06 | 91.40±0.40 | +5.94 | 92.23±0.26 | +6.77 | **92.48±0.20** | +7.02 |

**Adding MAKD.** We next apply MAKD to strengthen weaker modality representations before fusion and compare IAF+MAKD with Symmetric+MAKD under the same distillation setup (Table 5). IAF+MAKD outperforms Symmetric+MAKD in ten of eleven con-

figurations, even when both methods receive the same distillation signal. On MultiHuSE, the largest difference occurs with E5+VideoMAE: IAF+MAKD gains 2.06 pp over the unimodal ceiling, while Symmetric+MAKD falls 2.49 pp below it. IAF+MAKD is also higher on all five folds in three of four configurations (one-tailed Wilcoxon signed-rank, $p$=0.031); for E5+VideoMAE, it is higher on four of five folds ($p$=0.063).

On UR-FUNNY, IAF+MAKD achieves 69.80% versus 69.18% for Symmetric+MAKD and is higher on four of five seeds (Wilcoxon $p$=0.063; paired t-test $p$=0.032). On MUStARD, the differences are smaller: IAF+MAKD is higher on four of five folds for BERT+VideoMAE ($p$=0.063) and two of five for E5+VideoMAE ($p$=0.56). The only exception is Food-101 E5+PE-core, where Symmetric+MAKD performs better than IAF+MAKD (95.29% vs. 94.27%); on the other Food-101 configuration (BERT+PE-core), IAF+MAKD is slightly ahead (92.48% vs. 92.23%). All Food-101 methods improve by 5–9 pp over the unimodal ceiling, unlike the other datasets where fusion strategy has a larger effect. Additionally, we evaluate robustness to missing modalities at inference time in Appendix G.

**Pathway preservation.** Table 6 compares Symmetric+MAKD and IAF+MAKD unimodal pathway accuracies. IAF preserves the dominant pathway at its unimodal ceiling, whereas Symmetric+MAKD reduces it below the ceiling. Full results for all eleven configurations are in Appendix B. IAF is also competitive with prior methods on UR-FUNNY, MUStARD, and Food-101, while using fewer parameters and less computation than Symmetric Fusion (Appendix E, Appendix N).

Table 6: Pathway accuracy for Symmetric+MAKD and IAF+MAKD using distilled features; one representative configuration per dataset. Bold indicates pathways at their unimodal ceiling. Full results are in Appendix B.

| Dataset | Config. | Archi. | Ceiling | Text Path | Audio Path | Video/Image Path |
|---|---|---|---|---|---|---|
| MultiHuSE | E5+VideoMAE | Symmetric | 74.86 | 56.4 | 63.7 | 67.4 |
| | | IAF | 74.86 | **74.86** | 48.4 | 56.7 |
| Food-101 | BERT+PE-Core | Symmetric | 85.46 | 85.78 | – | 42.90 |
| | | IAF | 85.46 | 73.39 | – | **85.46** |
| UR-FUNNY | GloVe+OpenFace | Symmetric | 65.19 | 52.31 | 53.34 | 55.29 |
| | | IAF | 65.19 | **65.19** | 51.15 | 50.85 |
| MUStARD | BERT+PE-Core | Symmetric | 78.12 / 77.10 | 64.9 | 69.4 | 73.5 |
| | | IAF | 78.12 / 77.10 | 63.9 | **78.12** | **77.10** |

## 5 Conclusion

We identified *strong-modality collapse*, where multimodal fusion degrades the modality with the strongest unimodal performance. Causal decomposition shows that this degradation results from the interaction between cross-modal attention and classifier drift. *Inverted Asymmetric Fusion* (IAF) addresses this by removing cross-modal attention from the dominant pathway and freezing its classifier, while *Modality-Aware Knowledge Distillation* (MAKD) strengthens weaker modalities before fusion. Across four benchmarks with different dominant modalities, IAF improves over the strongest unimodal baseline, whereas the standard fusion strategies evaluated do not do so consistently. On the humor-style classification example from the introduction, early fusion falls 10.55 pp below the text-only model, while IAF+MAKD improves over the text-only baseline by 2.06 pp.

**Limitations.** IAF requires empirically identifying the dominant modality before fusion, which is less clear when modalities have similar predictive strength, and our main experiments use frozen features, with end-to-end fine-tuning tested only on a single MultiHuSE fold (see Appendix A for details).

### AI use statement

We used generative AI tools, including Claude and ChatGPT, to assist with writing and debugging experimental code and to proofread the manuscript and improve its clarity and

conciseness. We did not use generative AI to generate research ideas, conduct literature reviews, or produce experimental results; all reported numbers were obtained by executing our own code. The authors reviewed and verified all AI-assisted code and edited all AI-assisted text for accuracy. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with generative AI assistance.

### Ethics statement

This work uses four publicly available datasets (MultiHuSE, UR-FUNNY, MUStARD, Food-101), collected and released under standard academic licences by their respective authors. No new data collection or human annotation was conducted. All main experiments operate on pre-extracted, anonymised feature representations with no access to personally identifiable information.

### Reproducibility Statement

Full hyperparameter configurations, encoder details, and hardware setup are given in Appendix O. Computational cost is reported in Appendix N. MultiHuSE and MUStARD use 5-fold cross-validation, and UR-FUNNY and Food-101 use a five-seed robustness analysis (seeds 42, 100, 123, 600, 999). Code and pre-extracted feature embeddings for the datasets are available at an anonymous repository[2].

## References

Sajal Aggarwal, Ananya Pandey, and Dinesh Kumar Vishwakarma. Multimodal Sarcasm Recognition by Fusing Textual, Visual and Acoustic content via Multi-Headed Attention for Video Dataset. In *2023 World Conference on Communication and Computing, WCONF 2023*. Institute of Electrical and Electronics Engineers Inc., 2023. ISBN 9798350311204. doi: 10.1109/WCONF58270.2023.10235179.

Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, pp. 292–301. Association for Computing Machinery, Inc, 10 2018. ISBN 9781450356657. doi: 10.1145/3240508.3240578.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pp. 59–66. Institute of Electrical and Electronics Engineers Inc., 6 2018. ISBN 9781538623350. doi: 10.1109/FG.2018.00019.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy, 2 2019. ISSN 19393539.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception Encoder: The best visual embeddings are not at the output of the network. *arXiv Preprint (Facebook research)*, 4 2025. URL `http://arxiv.org/abs/2504.13181`.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, Florence, Italy, 7 2019. Association for Computational Linguistics. URL `https://github.`

[2]Code: `https://anonymous.4open.science/r/strong_modality_collaspe-3F67`

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A Closer Look at Multimodal Representation Collapse. In *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver, 8 2025. URL http://arxiv.org/abs/2505.22483.

Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4351–4360. Association for Computational Linguistics, 7 2020. URL http://www.iitp.ac.

Akshat Choube and Mohammad Soleymani. Punchline Detection using Context-Aware Hierarchical Multimodal Fusion. In *ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 675–679. Association for Computing Machinery, Inc, 10 2020. ISBN 9781450375818. doi: 10.1145/3382507.3418891.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. COVAREP: A Collaborative Voice Analysis Repository for Speech Technologies. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. IEEE, 2014. ISBN 9781479928934. URL http://covarep.github.io/covarep.

Boya Deng, Jianzhao Li, Maoguo Gong, Zedong Tang, Yourun Zhang, Kaiyuan Feng, and Yue Wu. BCFNet: Bi-temporal collaborative fusion network for multi-modal humor detection. *Pattern Recognition*, 172, 4 2025. ISSN 00313203. doi: 10.1016/j.patcog.2025.112744.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, Minnesota, 6 2019. Association for Computational Linguistics. URL https://github.com/tensorflow/tensor2tensor.

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang. Scaling up masked audio encoder learning for general audio classification. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 547–551. International Speech Communication Association, 2024. doi: 10.21437/Interspeech.2024-246.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical Modal Rebalance for Multimodal Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20038, Vancouver, 2023. IEEE. URL https://github.com/fanyunfeng-bit/Modal-Imbalance-PMR.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing Visual-linguistic Model via Knowledge Distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1408–1418. Institute of Electrical and Electronics Engineers Inc., 2021. ISBN 9781665428125. doi: 10.1109/ICCV48922.2021.00146.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 14755–14766. Association for Computational Linguistics, 11 2024. URL https://github.com/.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6):1789–1819, 6 2021. ISSN 15731405. doi: 10.1007/s11263-021-01453-z.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross Modal Distillation for Supervision Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2827–2836. IEEE, 2016.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131.

Association for Computing Machinery, Inc, 10 2020. ISBN 9781450379885. doi: 10.1145/3394171.3413678.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv Preprint*, 3 2015. URL http://arxiv.org/abs/1503.02531.

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 39th International Conference on Machine Learning*, Maryland, 2022.

Ignazio Gallo, Gianmarco Ria, Nicola Landro, and Riccardo La Grassa. Image and Text fusion for UPMC Food-101 using BERT and CNNs. In *International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pp. 1–6. IEEE, 2020. ISBN 9781728185798. doi: 10.1109/IVCNZ51579.2020.9290622.

Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. In *Proceedings of the 39th International Conference on Machine Learning,*, Maryland, 2022.

Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Hoque. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings ofthe 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2046–2056, 2019. URL www.ted.com.

Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. MultiHuSE: A Multimodal Dataset for Humour Styles and Emotions. In *International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–7. Institute of Electrical and Electronics Engineers (IEEE), 10 2025. doi: 10.1109/cbmi66578.2025.11339313.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, California, 7 2019. URL http://arxiv.org/abs/1905.00414.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting Multimodal Model Performance with Adaptive Gradient Modulation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22214–22224. Computer Vision Foundation, 2023a.

Songtao Li and Hao Tang. Multimodal Alignment and Fusion: A Survey. *International Journal of Computer Vision*, 10 2025. ISSN 15731405. doi: 10.1007/s11263-025-02667-1. URL http://arxiv.org/abs/2411.17040.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6631–6640. IEEE, 2023b. URL https://github.com/mdswyz/DMD.

Ronghao Lin and Haifeng Hu. Multi-Task Momentum Distillation for Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*, 15(2):549–565, 4 2024. ISSN 19493045. doi: 10.1109/TAFFC.2023.3282410.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2247–2256. Association for Computational Linguistics, 7 2018.

Pietro Morerio, Jacopo Cavazza, Riccardo Volpi, Rene Vidal, and Vittorio Murino. Curriculum Dropout. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pp. 3564–3572. Institute of Electrical and Electronics Engineers Inc., 12 2017. ISBN 9781538610329. doi: 10.1109/ICCV.2017.383.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, Chen Sun, and Google Research. Attention Bottlenecks for Multimodal Fusion. In *35th Conference on Neural Information Processing Systems*. NeurIPS, 2021.

Pradyumna Narayana, Aniket Pednekar, Abishek Krishnamoorthy, Kazoo Sone, and Sugato Basu. HUSE: Hierarchical Universal Semantic Embeddings. *arXiv*, 11 2019. URL http://arxiv.org/abs/1911.05978.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics, 2014. URL http://nlp.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3930–3940. IEEE, 2022.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, Ruslan Salakhutdinov, Louis-Philippe Morency, and Johns Hopkins. MULTIBENCH: Multiscale Benchmarks for Multimodal Representation Learning. In *35th Conference on Neural Information Processing Systems, Track on Datasets and Benchmarks.*, pp. 1–20. NeurIPS, 2021. URL https://cmu-multicomp-lab.github.io/multibench/.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369. Association for Computational Linguistics, 7 2020.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. A Multimodal Corpus for Emotion Recognition in Sarcasm. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pp. 20–25, 6 2022. URL https://github.com/apoorva-nunna/MUStARD_Plus_Plus.

Avantika Saklani, Shailendra Tiwari, and H. S. Pannu. Ameliorating multimodal food classification using state of the art deep learning techniques. *Multimedia Tools and Applications*, 83(21):60189–60212, 6 2024. ISSN 15737721. doi: 10.1007/s11042-023-17850-0.

Fida Mohammad Thoker and Juergen Gall. Cross-Modal Knowledge Distillation for Action Recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 9 2019. ISBN 9781538662496.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *36th Conference on Neural Information Processing Systems 9(NeurIPS)*, New Orleans, 11 2022. URL https://github.com/MCG-NJU/VideoMAE.

Yao Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 6558–6569. Association for Computational Linguistics (ACL), 2020. ISBN 9781950737482. doi: 10.18653/v1/p19-1656.

Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems*, 2017.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint*, 2 2024. URL http://arxiv.org/abs/2402.05672.

Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal Learning with Incomplete Modalities by Knowledge Distillation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1828–1838. Association for Computing Machinery, 8 2020a. ISBN 9781450379984. doi: 10.1145/3394486.3403234.

Tiannan Wang, Wangchunshu Zhou, Eth Zurich, Yan Zeng, and Xinsong Zhang. EfficientVLM: Fast and Accurate Vision-Language Models via Knowledge Distillation and Modal-adaptive Pruning. In *Findings of the Association for Computational Linguistics*, pp. 13899–13913. Association for Computational Linguistics, 7 2023. URL `https://github.com/swaggy-TN/EfficientVLM`.

Weiyao Wang, Du Tran, and Matt Feiszli. What Makes Training Multi-Modal Classification Networks Hard? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 12692–12702. IEEE Computer Society, 2020b. doi: 10.1109/CVPR42600.2020.01271.

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 and Expo Workshops, ICMEW 2015*. Institute of Electrical and Electronics Engineers Inc., 7 2015. ISBN 9781479970797. doi: 10.1109/ICMEW.2015.7169757.

Shicai Wei, Chunbo Luo, and Yang Luo. MMANet: Margin-aware Distillation and Modality-aware Regularization for Incomplete Multimodal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20039–20049. IEEE, 2023. URL `https://github.com/shicaiwei123/MMANet`.

Yake Wei and Di Hu. MMPareto: Boosting Multimodal Learning with Innocent Unimodal Assistance. In *Proceedings of the 41st International Conference on Machine Learning*, Vienna, 5 2024. URL `http://arxiv.org/abs/2405.17730`.

Yake Wei, Di Hu, Henghui Du, and Ji Rong Wen. On-the-Fly Modulation for Balanced Multimodal Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):469–485, 2025. ISSN 19393539. doi: 10.1109/TPAMI.2024.3468315.

Nan Wu, Stanisław Jastrzębski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. URL `https://github.com/nyukat/greedy_multimodal_learning`.

Yang Wu, Yanyan Zhao, Xin Lu, Bing Qin, Yin Wu, Jian Sheng, and Jinlong Li. Modeling Incongruity between Modalities for Multimodal Sarcasm Detection. *IEEE Multimedia*, 28 (2):86–95, 4 2021. ISSN 19410166. doi: 10.1109/MMUL.2021.3069097.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Denmark, 2017. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *The Sixth International Conference on Learning Representations (ICLR)*, Vancouver, 4 2018. URL `http://arxiv.org/abs/1710.09412`.

Yu Zhang, Bin Chen, Hongfei Ye, Zijian Gao, Tianjiao Wan, Long Lan, and Kele Xu. Text-guided Multimodal Fusion for the Multimodal Emotion and Intent Joint Understanding. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICASSP49660.2025.10890680.

Heqing Zou, Meng Shen, Chen Chen, Yuchen Hu, Deepu Rajan, and Eng Siong Chng. UniS-MMC: Multimodal Classification via Unimodality-supervised Multimodal Contrastive Learning. In *Findings of the Association for Computational Linguistics*, pp. 659–672. Association for Computational Linguistics, 7 2023.

# A Limitations

This work has several limitations. First, IAF requires identifying the dominant modality empirically through unimodal evaluation before fusion training. When modalities have similar predictive strength, the PURE/ATTEND assignment becomes less clear. The inferred hierarchy also depends on the encoder set: a stronger text encoder may appear dominant over a weaker video encoder even when the task relies more heavily on visual information. Thus, the hierarchy reflects empirical encoder performance rather than intrinsic task structure. Our dual-teacher design for MUStARD addresses co-dominant modalities, but more principled approaches, such as sample-level dynamic hierarchy assignment, remain open.

Second, all main experiments use frozen, pre-extracted features, isolating IAF's architectural contribution from encoder fine-tuning. We verify that its pathway-preservation property also holds with a live, updating encoder (Appendix F), but this check is limited to a single MultiHuSE fold with only the text encoder fine-tuned. Full end-to-end training across all folds, datasets, and modalities remains future work due to computational cost.

# B Full Pathway Preservation Results

Table 7 reports pathway accuracy for Symmetric+MAKD and IAF+MAKD across all eleven configurations, extending the representative cases in Table 6.

Table 7: Pathway accuracy for Symmetric+MAKD and IAF+MAKD. Each pathway is evaluated independently against its unimodal ceiling. Bold indicates pathways at their unimodal ceiling; MUStARD ceilings are Audio/Video. All metrics are %.

| Text | Video/Image Enc. | Archi. | Unimodal ceiling | Text Path | Audio Path | Video/Image Path |
|---|---|---|---|---|---|---|
| *MultiHuSE – Dominant Modality: TEXT – BERT=70.2%, E5=74.9%* | | | | | | |
| BERT | VideoMAE | Symmetric | 70.21±2.82 | 61.3±2.5 | 59.0±4.5 | 60.1±1.3 |
| | | IAF | 70.21±2.82 | **70.2±2.8** | 55.8±7.4 | 57.5±5.0 |
| | PE-Core | Symmetric | 70.21±2.82 | 63.5±1.6 | 63.6±2.3 | 61.6±7.2 |
| | | IAF | 70.21±2.82 | **70.2±2.8** | 62.9±2.5 | 65.6±3.1 |
| E5 | VideoMAE | Symmetric | 74.86±3.08 | 56.4±4.6 | 63.7±9.2 | 67.4±3.3 |
| | | IAF | 74.86±3.08 | **74.9±3.1** | 48.4±8.9 | 56.7±7.0 |
| | PE-Core | Symmetric | 74.86±3.08 | 55.3±8.7 | 65.5±2.2 | 69.7±4.8 |
| | | IAF | 74.86±3.08 | **74.9±3.1** | 54.9±9.6 | 62.2±6.6 |
| *Food101 – Dominant Modality: TEXT=86.24%; IMAGE (PE)=85.46%* | | | | | | |
| E5 | PE-Core | Symmetric | 86.24±0.11 | 78.98±9.33 | – | 74.72±8.62 |
| | | IAF | 86.24±0.11 | **86.24±0.11** | – | 84.60±0.16 |
| BERT | PE-Core | Symmetric | 85.46±0.06 | 85.78±2.70 | – | 42.89±9.39 |
| | | IAF | 85.46±0.06 | 73.39±0.83 | – | **85.46±0.06** |
| *UR-FUNNY – Dominant Modality: TEXT=65.19%* | | | | | | |
| GloVe | OpenFace | Symmetric | 65.19±0.09 | 52.31±2.90 | 53.34±3.53 | 55.29±5.35 |
| | | IAF | 65.19±0.09 | **65.19±0.09** | 51.15±1.29 | 50.85±0.28 |
| *MUStARD– Dominant Modalities: AUDIO=78.12%; VIDEO (PE)=77.10%* | | | | | | |
| BERT | VideoMAE | Symmetric | 78.12±5.45 / 74.49±1.16 | 62.0±5.0 | 64.4±15.2 | 67.8±8.0 |
| | | IAF | 78.12±5.45 / 74.49±1.16 | 64.6±9.0 | **78.1±5.5** | **74.5±1.2** |
| | PE-Core | Symmetric | 78.12±5.45 / 77.10±3.13 | 64.9±9.5 | 69.4±6.3 | 73.5±9.0 |
| | | IAF | 78.12±5.45 / 77.10±3.13 | 63.9±6.2 | **78.1±5.5** | **77.1±3.1** |
| E5 | VideoMAE | Symmetric | 78.12±5.45 / 74.49±1.16 | 60.3±12.3 | 77.3±5.1 | 71.5±6.2 |
| | | IAF | 78.12±5.45 / 74.49±1.16 | 66.4±7.6 | **78.1±5.5** | **74.5±1.2** |
| | PE-Core | Symmetric | 78.12±5.45 / 77.10±3.13 | 74.4±3.5 | 71.6±9.27 | 75.1±3.9 |
| | | IAF | 78.12±5.45 / 77.10±3.13 | 69.0±7.8 | **78.1±5.5** | **77.1±3.1** |

## C Full Unimodal Results

Table 8 reports base and MAKD-distilled unimodal accuracy and F1 for all modality/encoder combinations across the four benchmarks, extending the dominance hierarchy summarized in Table 1.

Table 8: Unimodal results across four benchmarks. *Base*: raw pretrained; *Dist.*: knowledge-distilled. ⋆ indicates the dominant modality. MultiHuSE and MUStARD: mean ± SD over 5-fold CV; UR-FUNNY and Food-101: mean ± SD over 5 seeds

| Modality / Feature | Base Acc(%) | Base F1(%) | Dist. Acc(%) | Dist. F1(%) |
|---|---|---|---|---|
| ***MultiHuSE*** | | | | |
| Text: BERT | 68.26±2.25 | 68 | 70.21±2.82 | 70 |
| Text: E5⋆ | **74.86±3.08** | **75** | — | — |
| Audio: Dasheng | 60.45±1.58 | 60 | 62.90±1.83 | 63 |
| Video: PE-Core | 43.66±1.29 | 43 | 45.57±1.22 | 46 |
| Video: VideoMAE | 40.13±0.73 | 40 | 41.01±1.60 | 41 |
| ***Food-101*** | | | | |
| Text: BERT | 76.08±0.03 | 76 | 75.3±0.06 | 75 |
| Text: E5⋆ | **86.24±0.11** | **86** | — | — |
| Image: PE-Core | 85.46±0.06 | 85 | 85.38±0.04 | 85 |
| ***UR-FUNNY*** | | | | |
| Text: GloVe⋆ | **62.11±1.22** | **64** | 65.19±0.09 | 65 |
| Audio: COVAREP | 57.55±2.13 | 48 | 63.78±1.04 | 64 |
| Video: OpenFace | 58.41±0.36 | 58 | 58.95±0.04 | 59 |
| ***MUStARD*** | | | | |
| Text: BERT | 69.42±2.80 | 69 | 72.03±2.49 | 72 |
| Text: E5 | 70.29±2.00 | 70 | 72.32±2.26 | 72 |
| Audio: Dasheng⋆ | **78.12±5.45** | **78** | — | — |
| Video: PE-Core⋆ | **77.10±3.13** | **77** | — | — |
| Video: VideoMAE | 74.49±1.16 | 74 | — | — |

## D Representational Similarity (CKA)

Table 9 reports linear CKA computed on raw, non-distilled features, matching the setting used in Section 4.3. CKA compares each fused pathway representation with its corresponding standalone unimodal representation; 1.0 indicates identical representations up to rotation and scaling. Bold rows indicate the dominant modality. MultiHuSE and MUStARD report mean ± SD over 5 folds; UR-FUNNY and Food-101 report mean ± SD over 5 seeds.

## E Comparison Against Prior Published Methods

Table 10, 11, and 12 compare our results with previously published methods on UR-FUNNY, MUStARD, and Food-101. MultiHuSE also has a published baseline (Kenneth et al., 2025), but its evaluation split is not grouped by text ID as in our experiments, making a direct comparison inappropriate; we therefore omit it. IAF achieves 69.80% on UR-FUNNY and 83.33% on MUStARD, while its Food-101 result is 94.27%, within 0.5 percentage points of the highest reported result shown in Table 12. These results provide context for the main experiments, showing that IAF's gains over the unimodal ceiling are accompanied by competitive overall performance.

## F Joint End-to-End Fine-Tuning (MultiHuSE)

All main results use frozen, pre-extracted features. We test whether IAF's dominant-pathway preservation holds when the dominant encoder is live and updating, using MultiHuSE (E5+PE-Core), fold 1 only, as a targeted robustness check rather than a full replica-

Table 9: Linear CKA between fused and standalone pathway representations across four benchmarks. *IAF vs. Uni* and *Sym. vs. Uni* compare each fused pathway with its standalone unimodal representation; $\Delta$ is the difference between them.

| Combination | Modality | IAF vs. Uni | Sym. vs. Uni | Diff (IAF−Sym) |
|---|---|---|---|---|
| | | *MultiHuSE* | | |
| BERT+VideoMAE | **Text** | **1.0000±0.0000** | **0.9429±0.0208** | **+0.0571** |
| | Audio | 0.9278±0.0108 | 0.9177±0.0053 | +0.0102 |
| | Video | 0.9128±0.0327 | 0.8742±0.0504 | +0.0386 |
| BERT+PE-Core | **Text** | **1.0000±0.0000** | **0.9265±0.0241** | **+0.0735** |
| | Audio | 0.9212±0.0130 | 0.9172±0.0158 | +0.0039 |
| | Video | 0.8786±0.0332 | 0.8226±0.0503 | +0.0560 |
| E5+VideoMAE | **Text** | **1.0000±0.0000** | **0.9097±0.1090** | **+0.0903** |
| | Audio | 0.9291±0.0182 | 0.9224±0.0154 | +0.0067 |
| | Video | 0.9209±0.0204 | 0.8636±0.0386 | +0.0572 |
| E5+PE-Core | **Text** | **1.0000±0.0000** | **0.9164±0.0888** | **+0.0836** |
| | Audio | 0.9240±0.0214 | 0.9169±0.0099 | +0.0071 |
| | Video | 0.8950±0.0470 | 0.8343±0.0344 | +0.0607 |
| | | *Food-101* | | |
| E5+PE-Core | **Text** | **1.0000 ± 0.0000** | **0.7467 ± 0.0438** | **+0.2533** |
| | Image | 0.9665 ± 0.0008 | 0.9539 ± 0.0048 | +0.0126 |
| BERT+PE-Core | Text | 0.8701 ± 0.0024 | 0.5672 ± 0.0091 | +0.3029 |
| | **Image** | **1.0000 ± 0.0000** | **0.9619 ± 0.0017** | **+0.0381** |
| | | *UR-FUNNY* | | |
| GloVe+COV+OF | **Text** | **1.0000 ± 0.0000** | **0.9426 ± 0.0022** | **+0.0574** |
| | Audio | 0.9305 ± 0.0050 | 0.9277 ± 0.0067 | +0.0028 |
| | Video | 0.9586 ± 0.0056 | 0.9519 ± 0.0046 | +0.0067 |
| | | *MUStARD* | | |
| BERT+VideoMAE | Text | 0.9698±0.0022 | 0.9688±0.0050 | +0.0010 |
| | **Audio** | **1.0000±0.0000** | **0.9304±0.0163** | **+0.0696** |
| | Video | 1.0000±0.0000 | 0.8531±0.0414 | +0.1469 |
| BERT+PE-Core | Text | 0.9618±0.0179 | 0.9127±0.1373 | +0.0491 |
| | **Audio** | **1.0000±0.0000** | **0.9284±0.0152** | **+0.0716** |
| | Video | 1.0000±0.0000 | 0.9215±0.0292 | +0.0785 |
| E5+VideoMAE | Text | 0.7770±0.2719 | 0.9626±0.0572 | -0.1856 |
| | **Audio** | **1.0000±0.0000** | **0.9216±0.0160** | **+0.0784** |
| | Video | 1.0000±0.0000 | 0.8236±0.0346 | +0.1764 |
| E5+PE-Core | Text | 0.8095±0.2667 | 0.6916±0.3930 | +0.1179 |
| | **Audio** | **1.0000±0.0000** | **0.9232±0.0102** | **+0.0768** |
| | Video | 1.0000±0.0000 | 0.9231±0.0247 | +0.0769 |

tion. Only the text encoder is fine-tuned via LoRA ($r$=8, $\alpha$=16, dropout 0.1); audio and video remain the trained, frozen MAKD-distilled backbones from fold 1. For computational efficiency, we use 20 rather than 100 epochs and warm-start the classifier heads. A full 5-fold joint fine-tuning run was estimated at ∼49 GPU-hours per configuration, whereas this reduced-scope check required ∼0.24 GPU-hours across the three architectures.

We compare three architectures, all using the same live, LoRA-updated text encoder and differing in fusion topology and whether the text (PURE) classifier head is frozen or trainable:

- **IAF, frozen head** – the paper's defined method: text bypasses cross-modal attention, and its classifier is frozen and warm-started from the frozen-feature checkpoint.

Table 10: UR-FUNNY: comparison against prior published methods. Context: ✓ = punchline + context; × = punchline only. †Mean over 5 seeds.

| Method | Context | Acc (%) | F1 (%) |
|---|---|---|---|
| C-MFN (Kamrul Hasan et al., 2019) | × | 64.47 | – |
| C-MFN (Kamrul Hasan et al., 2019) | ✓ | 65.23 | – |
| AGM(Late fusion) (Li et al., 2023a) | × | 65.97 | – |
| AGM(Early fusion) (Li et al., 2023a) | × | 66.07 | – |
| MULTIBENCH (Pu Liang et al., 2021) | ✓ | 66.70 | – |
| BCFNet (Deng et al., 2025) | × | 66.78 | 68 |
| HF (Choube & Soleymani, 2020) | ✓ | 67.84 | 69 |
| MISA (GloVe) (Hazarika et al., 2020) | × | 68.60 | – |
| **Proposed (IAF)** | ✓ | **69.80**† | 71 |

Table 11: MUStARD: comparison against prior published methods.

| Method | Context | Acc(%) | F1(%) |
|---|---|---|---|
| MULTIBENCH (Pu Liang et al., 2021) | ✓ | 71.8±0.3 | — |
| SVM (BERT) (Castro et al., 2019) | ✓ | — | 71.5 |
| SWIA (Chauhan et al., 2020) | ✓ | — | 72.6 |
| CE (BART) (Ray et al., 2022) | ✓ | — | 74.2 |
| MuLOT (BERT) (Pramanick et al., 2022) | × | 74.52 | — |
| IWAN (BERT) (Wu et al., 2021) | × | — | 74.5 |
| IWAN (BERT) (Wu et al., 2021) | ✓ | — | 75.1 |
| MuLOT (BERT) (Pramanick et al., 2022) | ✓ | 76.82 | — |
| MHA (RoBERTa) (Aggarwal et al., 2023) | ✓ | 79.32 | 77.6 |
| **Proposed (IAF (E5))** | ✓ | **83.33 ±4.20** | **84.0** |

- **IAF, trainable head** – an exploratory variant: text still bypasses cross-modal attention, but its classifier is trainable.
- **Symmetric Fusion** – all three modalities cross-attend to one another: the text projection and shared classifier are trainable. No frozen-head variant is defined because Symmetric Fusion has no PURE pathway.

Table 13 reports text-pathway accuracy and full fused accuracy for each architecture, relative to the fold-1 unimodal E5 ceiling of 73.24% obtained with frozen features.

The text-pathway results show a smaller drop under IAF than under Symmetric Fusion: −2.70 pp with a frozen head and −4.77 pp with a trainable head, compared with −9.54 pp for Symmetric Fusion. This supports the pathway-preservation property of IAF when the dominant encoder is jointly fine-tuned. Freezing the PURE head also reduces the pathway drop relative to the trainable-head variant, consistent with the causal decomposition, which identifies classifier adaptation as one component of the degradation. The gate assigns text the highest average weight (0.56) even in the trainable-head variant, indicating that the dominant pathway continues to receive substantial gradient signal in this configuration.

## G Missing-Modality Robustness

Pathway analysis and causal decomposition evaluate representations with all modalities present. We therefore test whether the full gated model remains robust when a modality is absent at inference. We zero one modality's backbone feature and pass the resulting inputs through the complete model on Food-101 and MultiHuSE (Tables 14 and 15).

On Food-101 (Table 14), IAF matches or exceeds Symmetric Fusion in 5 of 6 comparisons across the two encoder configurations and three masking conditions. The only exception

Table 12: Comparison against prior published methods on UPMC Food-101, using the same standard 67,972/22,716 train/test split and 101-class taxonomy as this paper.

| Method | Image | Text | Fusion |
|---|---|---|---|
| Wang et al. (2015) | 40.2 | 82.0 | 85.1 |
| Narayana et al. (2019) | 73.8 | 87.3 | 92.3 |
| Ignazio Gallo et al. (2020) | 71.67 | 84.6 | 92.5 |
| Zou et al. (2023) | – | – | 94.7 |
| Saklani et al. (2024) | 70.59 | 92.49 | 94.06 |
| IAF (this paper, E5+PE-Core) | 85.46±0.06 | 86.24±0.11 | 94.27 ± 0.35 |
| IAF (this paper, BERT+PE-Core) | 85.46±0.06 | 76.08±0.03 | 92.48 ± 0.20 |

Table 13: Joint end-to-end fine-tuning on MultiHuSE fold 1 (E5+PE-Core). Text-pathway accuracy is measured using the text/PURE classifier after training; fused accuracy is the accuracy of the full gated model. The unimodal ceiling is 73.24%, from the fold-1 E5 unimodal model with frozen features.

| Architecture | Text pathway acc. (%) | Δ vs. ceiling | Fused acc. (%) |
|---|---|---|---|
| IAF, frozen head | 70.54 | −2.70 | 73.03 |
| IAF, trainable head | 68.46 | −4.77 | 74.07 |
| Symmetric fusion | 63.69 | −9.54 | 69.29 |

occurs for BERT+PE-Core when the dominant image modality is masked, where Symmetric Fusion achieves 76.39% versus 73.87% for IAF.

MultiHuSE shows the same pattern more consistently (Table 15). Across all four configurations, IAF matches or exceeds Symmetric Fusion when at least one non-dominant modality remains available: full input and masking of either audio or video give 12 of 12 comparisons in favor of or tied with IAF. When the dominant text modality is fully masked, Symmetric Fusion performs better in all four configurations, with differences of 0.25–1.50 pp in three configurations and 3.62 pp for E5+VideoMAE.

Table 14: Missing-modality robustness, Food-101 (Seed 999): accuracy (%) and gate weights under full input and single-modality masking. $w_T, w_I$ = mean gate weight on text and image.

| Config. | Architecture | Scenario | Acc. (%) | $(w_T, w_I)$ |
|---|---|---|---|---|
| E5+PE-Core (text dom.) | Symmetric | Full | 93.93 | (0.560, 0.440) |
| | | Image-masked | 86.00 | (0.010, 0.990) |
| | | Text-masked | 84.21 | (0.656, 0.344) |
| | IAF | Full | 94.43 | (0.310, 0.690) |
| | | Image-masked | 86.31 | (0.660, 0.340) |
| | | Text-masked | 85.20 | (0.002, 0.998) |
| BERT+PE-Core (image dom.) | Symmetric | Full | 90.36 | (0.779, 0.221) |
| | | Image-masked | 76.39 | (0.001, 0.999) |
| | | Text-masked | 85.10 | (0.828, 0.172) |
| | IAF | Full | 91.03 | (0.517, 0.483) |
| | | Image-masked | 73.87 | (0.999, 0.001) |
| | | Text-masked | 85.40 | (0.318, 0.682) |

The gate behavior is consistent with these results. Under IAF, masking a modality shifts weight toward the surviving modality in every Food-101 condition where this comparison is available; for example, when image is masked in E5+PE-Core, the image weight falls from 0.690 to 0.340 while the text weight rises from 0.310 to 0.660. In contrast, Symmetric Fusion can assign increased weight to a zeroed modality: when image is masked in E5+PE-

Table 15: Missing-modality robustness, MultiHuSE (distilled, 5-fold mean): accuracy (%) under full input and single-modality masking. T = text (dominant), A = audio, V = video.

| Config. | Arch. | Full (T+A+V) | T+A (V missing) | T+V (A missing) | A+V (T missing) |
|---|---|---|---|---|---|
| bert+videomae | Symmetric | 69.42 | 67.22 | 61.03 | **60.99** |
| bert+videomae | IAF | 73.16 | 70.25 | 65.85 | **59.49** |
| bert+PE-Core | Symmetric | 72.00 | 68.26 | 66.64 | **63.98** |
| bert+PE-Core | IAF | 74.74 | 70.34 | 69.50 | **63.73** |
| e5+videomae | Symmetric | 72.37 | 70.59 | 67.72 | **62.11** |
| e5+videomae | IAF | 76.92 | 73.78 | 72.20 | **58.49** |
| e5+PE-Core | Symmetric | 74.32 | 71.46 | 72.74 | **63.11** |
| e5+PE-Core | IAF | 78.06 | 75.24 | 73.28 | **62.86** |

Core, its image weight rises from 0.440 to 0.990. We hypothesize that this behavior reflects the curriculum modality-dropout schedule used during training (Appendix K). Overall, IAF remains competitive when a non-dominant modality is missing, while removing the dominant modality produces the largest relative disadvantage for IAF, consistent with its protected dominant-pathway design.

## H Role of Modalities Ablation

We remove one modality at a time to quantify its contribution to full trimodal performance. Table 16 shows that full trimodal IAF outperforms every bimodal variant on all three datasets. On MultiHuSE, removing text reduces accuracy from 78.06% to 65.85%, consistent with text's dominant role. On UR-FUNNY, $T{+}A$ outperforms $T{+}V$ (68.91% vs. 65.29%), consistent with the unimodal hierarchy. On MUStARD, removing text has a smaller effect: the text-free $A{+}V$ pair achieves 82.61% compared with 83.33% for the full model, consistent with the co-dominance of audio and video.

Table 16: Bimodal ablation for the E5- and GloVe-based encoder configurations.

| **Configuration** | **Pure** | **Attends** | **Acc(%) ± SD** | **F1(%)** |
|---|---|---|---|---|
| *MultiHuSE — Text dominant (E5, Dasheng, PE-Core)* | | | | |
| Full Trimodal | Text | Audio, Video | **78.06 ±3.82** | **78** |
| T+V | Text | Video | 76.61 ±4.58 | 77 |
| T+A | Text | Audio | 76.57 ±2.92 | 76 |
| A+V (Text-Free) | Audio | Video | 65.85 ±1.63 | 65 |
| *UR-FUNNY — Text dominant (GloVe, COVAREP, OpenFace)* | | | | |
| Full Trimodal | Text | Audio, Video | **69.80** | **70** |
| T+A | Text | Audio | 68.91 | 69 |
| T+V | Text | Video | 65.29 | 65 |
| A+V (Text-Free) | Audio | Video | 65.29 | 65 |
| *MUStARD — Audio/Video dominant (E5, Dasheng, PE-Core)* | | | | |
| Full Trimodal | Audio, Video | Text | **83.33 ±4.20** | **84** |
| A+T | Audio | Text | 79.57 ±5.17 | 80 |
| V+T | Video | Text | 79.57 ±5.41 | 80 |
| A+V (Text-Free) | Audio | Video | 82.61 ±3.01 | 83 |

Table 17 extends the ablation to BERT encoder configurations on MultiHuSE and MUStARD. The results show the same pattern: full trimodal IAF outperforms the reported bimodal variants.

Table 17: Bimodal fusion for BERT configurations on MultiHuSE and MUStARD.

| Configuration | Pure | Attends | Acc (%) ± SD | F1(%) |
|---|---|---|---|---|
| *MultiHuSE — Text dominant (BERT, Dasheng, PE-Core)* | | | | |
| Full Trimodal | Text | Audio, Video | **74.74 ± 1.71** | **74** |
| T+V | Text | Video | 72.45 ± 3.27 | 72 |
| T+A | Text | Audio | 72.08 ± 2.71 | 72 |
| *MUStARD — Audio/Video dominant (BERT, Dasheng, PE-Core)* | | | | |
| Full Trimodal | Audio, Video | Text | **82.61 ± 3.92** | **83** |
| A+T | Audio | Text | 79.71 ± 3.30 | 80 |
| V+T | Video | Text | 77.25 ± 4.75 | 77 |

## I Gate Mechanism Ablation

Table 18 compares three gating strategies within IAF: a *Learnable* gate with sample-adaptive weights, a *Hierarchy* gate with fixed weights derived from the modality ranking, and a *Uniform* gate with equal weights. The Learnable gate performs best on UR-FUNNY and MUStARD, while the Hierarchy gate performs slightly better on MultiHuSE. The differences are small, ranging from 0.04 to 1.45 percentage points across the evaluated configurations, indicating that IAF's performance is not strongly dependent on the choice of gating strategy. The mean learned gate weights also remain distributed across modalities, with no modality receiving more than 0.53 weight on average.

Table 18: Gate mechanism ablation across three datasets. Learnable = sample-adaptive weights; Hierarchy = fixed weights derived from modality ranking; Uniform = equal weights (T=A=V=0.33). Δ = accuracy difference relative to the Learnable gate.

| Config | Gate Type | Acc(%) ± SD | Δ(%) |
|---|---|---|---|
| *MultiHuSE Dataset* (Hierarchy: T=0.50, A=0.30, V=0.20) | | | |
| BERT + PE-Core | Learnable | 74.74 ±1.71 | – |
| | Hierarchy | **74.99 ±2.68** | +0.25 |
| | Uniform | 74.86 ±3.31 | +0.12 |
| E5 + PE-Core | Learnable | 78.06 ±3.82 | – |
| | Hierarchy | **78.52 ±3.64** | +0.46 |
| | Uniform | 78.02 ±3.81 | −0.04 |
| *UR-FUNNY Dataset: Seed–999* (Hierarchy: T=0.50, A=0.30, V=0.20) | | | |
| GloVe/COVAREP/OpenFace | Learnable | **70.72** | – |
| | Hierarchy | 70.32 | −0.40 |
| | Uniform | 69.42 | −1.30 |
| *MUStARD Dataset* (Hierarchy: T=0.25, A=0.40, V=0.35) | | | |
| BERT + PE-Core | Learnable | **82.61 ±3.92** | – |
| | Hierarchy | 81.74 ±4.11 | −0.87 |
| | Uniform | 81.88 ±3.94 | −0.73 |
| E5 + PE-Core | Learnable | **83.33 ±4.20** | – |
| | Hierarchy | 82.03 ±4.48 | −1.30 |
| | Uniform | 81.88 ±4.42 | −1.45 |

## J Gate Weight Analysis

The IAF gate produces per-sample weights $(w_T, w_A, w_V)$ that sum to one and determine each modality's contribution to the final prediction. Table 19 reports mean gate weights at the best checkpoint across the evaluated configurations. The weights generally reflect

the PURE/ATTEND assignments: text receives the highest mean weight in three of four MultiHuSE configurations, while audio and video together receive more than 0.80 of the weight on all MUStARD configurations, where both are PURE. UR-FUNNY has a more balanced distribution across modalities ($\bar{w}_T = 0.372$, $\bar{w}_A = 0.310$, $\bar{w}_V = 0.318$). No modality exceeds a mean weight of 0.53 in any configuration, indicating that the learned gate does not assign near-exclusive average weight to a single modality.

Table 19: Mean learned gate weights at the best checkpoint. MultiHuSE and MUStARD are averaged across CV folds; UR-FUNNY uses seed 999. $\Delta$ = gain over the dominant unimodal baseline (MultiHuSE: BERT = 68.26%, E5 = 74.86%; UR-FUNNY: GloVe = 62.11%; MUStARD: Dasheng = 78.12%). $^{[B]}/^{[E]}$: gap from BERT/E5 baseline. Base = raw pre-trained; Dist. = knowledge-distilled.

| **Text** | **Feat.** | **$\Delta$(%)** | $\bar{w}_T$ | $\bar{w}_A$ | $\bar{w}_V$ |
|---|---|---|---|---|---|
| **MultiHuSE** *(Text PURE \| Dasheng+PE-Core)* | | | | | |
| BERT+PE-Core | Base | $+4.07^{[B]}$ | 0.474 | 0.355 | 0.171 |
| | Dist. | $+6.48^{[B]}$ | 0.355 | 0.206 | 0.439 |
| E5+PE-Core | Base | $+1.04^{[E]}$ | 0.431 | 0.425 | 0.144 |
| | Dist. | $+3.20^{[E]}$ | 0.412 | 0.197 | 0.391 |
| **UR-FUNNY** *(Text PURE \| COVAREP + OpenFace)* | | | | | |
| GloVe | Dist. | +7.69 | 0.372 | 0.310 | 0.318 |
| **MUStARD** *(Audio+Video PURE \| Text ATTEND)* | | | | | |
| BERT+PE-Core | Base | $+2.02^{[B]}$ | 0.194 | 0.530 | 0.276 |
| | Dist. | $+4.49^{[B]}$ | 0.161 | 0.389 | 0.449 |
| E5+PE-Core | Base | $+3.62^{[E]}$ | 0.166 | 0.440 | 0.394 |
| | Dist. | $+5.21^{[E]}$ | 0.103 | 0.470 | 0.428 |

Mean weights can mask substantial sample-level variation. Table 20 shows representative examples from the distilled MultiHuSE E5+PE-Core model. Although the mean text weight is approximately 0.41, individual samples range from near-exclusive text weighting ($w_T > 0.95$) to strong audio weighting ($w_A > 0.85$). Samples with high text weights can be classified correctly, while others are misclassified despite assigning most weight to text. Conversely, several correctly classified samples assign more than 0.85 weight to audio despite text being the empirically dominant modality. These examples illustrate that the gate adapts its weighting at the sample level rather than applying the population-level modality hierarchy uniformly.

## K Regularization Ablation

**Mixup.** For each mini-batch, we sample a mixing coefficient $\lambda \sim \text{Beta}(\alpha_{\text{mix}}, \alpha_{\text{mix}})$ with $\alpha_{\text{mix}}$=0.05 and interpolate pairs of feature vectors and their labels:

$$\tilde{\mathbf{h}}_i^{(m)} = \lambda\, \mathbf{h}_i^{(m)} + (1-\lambda)\, \mathbf{h}_{\sigma(i)}^{(m)}, \quad \tilde{y}_i = (\lambda,\, y_i,\, y_{\sigma(i)}) \tag{9}$$

where $\sigma$ is a random permutation of batch indices. Mixup regularizes training by encouraging predictions that vary smoothly between training examples.

**Curriculum Modality Dropout.** To reduce reliance on any single modality, we apply a two-phase masking schedule. All modalities are present during warm-up (epochs <15). Thereafter, each sample is assigned one of seven masking configurations: trimodal ($p$=0.40), each of the three bimodal pairs ($p$=0.10), or each of the three unimodal cases ($p$=0.10). Masked modalities are zeroed before the forward pass.

Table 20: Representative IAF gate weight profiles (E5 + PE-Core, MultiHuSE, all folds). *This table contains potentially offensive or sensitive example text reproduced from the dataset for analysis.* Class labels: Aff = Affiliative, Agg = Aggressive, Neu = Neutral, S-enh = Self-enhancing. $w_T+w_A+w_V = 1$ per sample. ✓ correct; ✗ incorrect.

| **Sample text** | **True** | **Pred** | $w_T$ | $w_A$ | $w_V$ | |
|---|---|---|---|---|---|---|
| *(i) High text weight, correctly classified* | | | | | | |
| Joe Kennedy III reveals how his GOP counterparts really feel about Trump's tweets. | Neu | Neu | 0.970 | 0.006 | 0.024 | ✓ |
| Katy Perry wears American flag dress to kids in overall concerts. | Neu | Neu | 0.961 | 0.018 | 0.020 | ✓ |
| Porn actress confirms Trump affair after unpublished 2011 interview. | Neu | Neu | 0.959 | 0.020 | 0.021 | ✓ |
| *(ii) Low text weight, correctly classified* | | | | | | |
| The only positive thing about you is your HIV status. | Agg | Agg | 0.023 | 0.548 | 0.429 | ✓ |
| Advice to ice skaters: you can't always tell a brick by its cover. | Neu | Neu | 0.024 | 0.643 | 0.333 | ✓ |
| He who laughs last probably doesn't understand the joke. | Neu | Neu | 0.028 | 0.685 | 0.287 | ✓ |
| *(iii) High text weight, incorrectly classified* | | | | | | |
| Age is of no importance unless you're a cheese. | S-enh | Agg | 0.885 | 0.032 | 0.082 | ✗ |
| Don't look back. You're not going that way. | S-enh | Agg | 0.863 | 0.049 | 0.088 | ✗ |
| What is the best contraceptive for old people? Nudity. | Agg | Aff | 0.857 | 0.084 | 0.058 | ✗ |
| *(iv) Audio-dominant, correctly classified* | | | | | | |
| How many Californians does it take to screw in a light bulb? None. They screw in hot tubs. | Neu | Neu | 0.039 | 0.900 | 0.061 | ✓ |
| Tomorrow is 2-22-22. Happy Tuesday! | Aff | Aff | 0.135 | 0.856 | 0.009 | ✓ |
| Why couldn't the leopard play hide and seek? Because he was always spotted. | Aff | Aff | 0.102 | 0.851 | 0.048 | ✓ |

Table 21 reports the effect of removing each regularizer individually and both together. Removing both produces the largest drop in every evaluated configuration, ranging from 1.61 to 4.07 pp, while removing either regularizer alone produces smaller decreases. The largest drop occurs on MultiHuSE with E5+Dasheng+PE-Core (4.07 pp).

## L Distillation Coefficient Sensitivity

Table 22 reports the effect of varying the three distillation loss coefficients ($\alpha$: hard CE, $\beta$: soft KD, $\gamma$: feature alignment) on MultiHuSE. We compare the default setting with three alternatives: equal weights, a CE-heavy variant, and a soft-target-heavy variant. The distilled students are then used as inputs to IAF.

IAF fusion accuracy varies by at most 1.20 percentage points across the four settings. The default and equal-weight settings both achieve 78.06%, while the CE-heavy and soft-target-heavy settings achieve 76.86% and 76.90%, respectively. The CE-heavy setting gives the highest video distillation accuracy (45.95%), a 0.38 pp increase over the default, but its overall fusion accuracy is 1.20 pp lower.

The soft-target-heavy setting reduces video distillation accuracy by 0.83 pp relative to the default, while audio distillation remains stable, with a range of only 0.33 pp across settings. Overall, the results indicate that IAF is not highly sensitive to the precise coefficient values. We use the default setting ($(\alpha$=0.40,$\beta$=0.35,$\gamma$=0.25)) throughout the main experiments; these coefficients were selected by manual tuning.

## M Knowledge Distillation Design Comparison

Table 23 compares MAKD with prior multimodal distillation methods across four design axes: training stage, teacher source, purpose, and optimization procedure. MAKD is applied as a standalone pre-fusion stage, using the empirically dominant unimodal modality

Table 21: Regularization ablation across MultiHuSE, UR-FUNNY, and MUStARD. ✓✓ = both Mixup and Curriculum Dropout enabled (full model); Δ = accuracy change relative to the full model. All metrics in %.

| Mixup | Curriculum | Acc (%) ± SD | F1 | Δ (%) |
|---|---|---|---|---|
| **MultiHuSE** *(BERT + Dasheng + PE-Core)* | | | | |
| ✓ | ✓ | **74.74 ±1.71** | **74** | – |
| ✗ | ✓ | 73.49 ±3.15 | 73 | −1.25 |
| ✓ | ✗ | 72.58 ±3.80 | 72 | −2.16 |
| ✗ | ✗ | 70.75 ±2.67 | 70 | −3.99 |
| **MultiHuSE** *(E5 + Dasheng + PE-Core)* | | | | |
| ✓ | ✓ | **78.06 ±3.82** | **78** | – |
| ✗ | ✓ | 76.61 ±4.40 | 77 | −1.45 |
| ✓ | ✗ | 75.74 ±5.53 | 76 | −2.32 |
| ✗ | ✗ | 73.99 ±5.04 | 74 | −4.07 |
| **UR-FUNNY: Seed–999** *(GloVe + COVAREP + OpenFace)* | | | | |
| ✓ | ✓ | **70.72** | **71** | – |
| ✓ | ✗ | 70.62 | 71 | −0.10 |
| ✗ | ✓ | 69.72 | 70 | −1.00 |
| ✗ | ✗ | 69.11 | 69 | −1.61 |
| **MUStARD** *(BERT + Dasheng + PE-Core)* | | | | |
| ✓ | ✓ | **82.61 ±3.92** | **83** | – |
| ✓ | ✗ | 82.03 ±4.08 | 82 | −0.58 |
| ✗ | ✓ | 81.45 ±4.11 | 81 | −1.16 |
| ✗ | ✗ | 79.71 ±4.25 | 80 | −2.90 |
| **MUStARD** *(E5 + Dasheng + PE-Core)* | | | | |
| ✓ | ✓ | **83.33 ±4.20** | **84** | – |
| ✓ | ✗ | 82.03 ±3.73 | 82 | −1.30 |
| ✗ | ✓ | 81.59 ±3.65 | 82 | −1.74 |
| ✗ | ✗ | 81.16 ±4.51 | 81 | −2.17 |

Table 22: Distillation coefficient sensitivity on MultiHuSE (5-fold CV, mean ± SD %). † = setting used in all main experiments. IAF use distilled students; Audio = Dasheng, Video = PE-Core throughout. "Range" reports max − min across the four settings.

| Setting | Coeff. ($\alpha/\beta/\gamma$) | BERT Dist. | Audio Dist. | Video Dist. | IAF (BERT) | IAF (E5) |
|---|---|---|---|---|---|---|
| Default† | 0.40 / 0.35 / 0.25 | **70.21**±2.82 | 62.90±1.83 | 45.57±1.22 | **74.74**±1.71 | **78.06**±3.82 |
| Equal weights | 0.33 / 0.33 / 0.33 | 69.92±2.79 | **63.11**±1.54 | 44.66±1.57 | 74.37±2.33 | **78.06**±3.36 |
| CE-heavy | 0.60 / 0.25 / 0.15 | 69.63±2.75 | 62.78±1.18 | **45.95**±0.98 | 74.03±2.48 | 76.86±4.32 |
| Soft-target-heavy | 0.25 / 0.55 / 0.20 | 70.17±2.58 | 62.86±1.10 | 44.74±1.35 | 73.33±2.32 | 76.90±4.18 |
| Range | – | 0.58 | 0.33 | 1.29 | 1.41 | 1.20 |

as teacher to strengthen weaker encoders before fusion. In contrast, the other compared methods primarily perform distillation within the fusion stage.

## N Computational Cost

Because the backbone encoders are frozen, the measured cost isolates the fusion stage. Table 24 reports trainable parameters, FLOPs per sample, training time per epoch, and inference latency for three-modality MultiHuSE and two-modality Food-101. On MultiHuSE, IAF uses 26.6% fewer trainable parameters and 28.5% fewer FLOPs per sample than Symmetric Fusion, reflecting the removal of one cross-attention block. On Food-101, the reductions are 39.2% and 37.3%, respectively. IAF also has lower epoch time and inference latency in both datasets across the configurations tested.

## O Full Hyperparameter Configurations

**Infrastructure and Evaluation:** Feature extraction is performed on an HPC cluster with an NVIDIA A16 (15,GB VRAM). All subsequent training is conducted on a consumer

Table 23: Design comparison of multimodal knowledge distillation methods.

| Method | Stage | Teacher | Purpose | Procedure |
|---|---|---|---|---|
| Cross-M KD Albanie et al. (2018) Thoker & Gall (2019) | Pre-fusion | Paired modality | Unimodal transfer | Sequential |
| Wang et al. Wang et al. (2020a) | Within fusion | Full multimodal | Dropout robustness | Joint |
| MMANet Wei et al. (2023) | Within fusion | Full multimodal | Missing-modality | Joint |
| DMD Li et al. (2023b) | Within fusion | Subspace teachers | Repr. disentanglement | Concurrent |
| MTMD Lin & Hu (2024) | Within fusion | Unimodal + fused | Heterogeneity reduction | Concurrent |
| **MAKD (Ours)** | **Pre-fusion** | **Dom. unimodal** | **Encoder strengthening** | **Sequential** |

Table 24: Fusion-stage computational cost. Backbone encoders are frozen and identical across architectures. MultiHuSE times represent the range across 4 encoder configurations and 5 folds; Food-101 times represent the range across two encoder combinations.

| Dataset | Architecture | Params | FLOPs/sample | Epoch (s) | Infer. (ms) |
|---|---|---|---|---|---|
| MultiHuSE | Symmetric | 3,947,016 | 5,523,456 | 1.053–1.305 | 0.213–0.226 |
| | IAF | **2,897,933** | **3,951,616** | **0.898–0.963** | **0.156–0.193** |
| Food-101 | Symmetric | 2,681,959 | 2,682,368 | 2.72–3.06 | 0.0266–0.0268 |
| | IAF | **1,630,311** | **1,682,432** | **2.29–2.33** | **0.0251–0.0252** |

laptop with an NVIDIA RTX 3050 (4,GB VRAM) using PyTorch. Unimodal pre-evaluation takes under 15 minutes per encoder on our hardware. MultiHuSE and MUStARD use 5-fold cross-validation, while UR-FUNNY and Food-101 use five seeds (42, 100, 123, 600, 999).

**Encoders.** Table 25 lists the dimensionality and parameter count of each pretrained encoder. All encoders remain frozen; no encoder parameters are updated during distillation or fusion training.

Table 25: Pretrained encoder dimensionality and parameter counts.

| Modality | Encoder | Dim. | Params |
|---|---|---|---|
| Text | BERT-base (Devlin et al., 2019) | 768 | 110M |
| Text | E5-large-instruct (Wang et al., 2024) | 1024 | 560M |
| Audio | Dasheng-0.6B (Dinkel et al., 2024) | 1280 | 600M |
| Video | VideoMAE (Tong et al., 2022) | 768 | 94M |
| Image/Video | PE-Core (Bolya et al., 2025) | 1024 | 320M (visual) / 310M (language) |

Table 27 and 26 report the complete hyperparameter configurations for the distillation and fusion stages. Hyperparameters were selected through manual tuning based on validation performance and then fixed for the reported experiments. No automated hyperparameter search was used.

Table 26: Dataset-specific hyperparameters for the fusion stage.

| Hyperparameter | MultiHuSE | UR-FUNNY | MUStARD | Food-101 |
|---|---|---|---|---|
| Optimizer | AdamW | Adam | AdamW | Adam |
| Learning rate | $6 \times 10^{-5}$ | $1 \times 10^{-5}$ | $1 \times 10^{-3}$ | $5 \times 10^{-5}$ |
| Weight decay | $10^{-3}$ | $10^{-3}$ | $10^{-2}$ | $10^{-3}$ |
| Max epochs | 100 | 80 | 100 | 40 |
| Batch size | 64 | 64 | 64 | 256 |
| Curriculum dropout warm-up | 15 epochs | 15 epochs | 15 epochs | 8 epochs |
| Early stopping patience | 35 | 20 | 40 | 10 |
| Gradient clipping | 1.0 | 1.0 | 1.0 | 1.0 |
| Mixup $\alpha_{\text{mix}}$ | 0.05 (all datasets) | | | |
| Cross-Modal Anchor $d_{\text{model}}$ | 512 (all datasets) | | | |
| Cross-Modal Anchor heads $H$ | 8 (all datasets) | | | |

Table 27: Distillation stage hyperparameters. Learning rates are modality- or experiment-specific as shown.

| **Hyperparameter** | **Value** |
|---|---|
| KD loss weights $(\alpha, \beta, \gamma)$ | 0.4, 0.35, 0.25 |
| Temperature $T$ | 3.5 |
| Optimizer | AdamW |
| Weight decay | $10^{-3}$ |
| *MultiHuSE* | |
| Text student LR | $1 \times 10^{-3}$ |
| Audio student LR | $1 \times 10^{-4}$ |
| Video student LR | $9 \times 10^{-5}$ |
| *UR-FUNNY* | |
| Text student LR | $1 \times 10^{-4}$ |
| Audio student LR | $1 \times 10^{-3}$ |
| Video student LR | $1.5 \times 10^{-3}$ |
| *MUStARD* | |
| BERT student LR | $1 \times 10^{-4}$ |
| E5 student LR (audio + video) | $5 \times 10^{-3}$ |
| *Food-101* | |
| Image student LR | $1 \times 10^{-4}$ |
| Text student LR | $1 \times 10^{-3}$ |