# Mitigating Strong-Modality Collapse in Multimodal Learning via Inverted Asymmetric Fusion

**Anonymous ACL submission**

## Abstract

Fusing multiple modalities is expected to improve model performance. However, on the MultiHuSE dataset, early, late, and symmetric attention fusion often fail to outperform the best unimodal baseline (text). Pathway isolation of a symmetric attention fusion model reveals that the text-pathway accuracy drops from 74.9% to 56.4% after fusion in one such setting, indicating that the dominant modality can be degraded during integration. We term this strong-modality collapse and argue that it helps explain why some multimodal models fail to surpass unimodal baselines. We propose Inverted Asymmetric Fusion (IAF), which avoids forcing mutual attention across modalities. The dominant modality is preserved by passing through fusion unchanged, while weaker modalities attend to it as a contextual anchor. Before fusion, weaker modalities are strengthened using Modality-Aware Knowledge Distillation. We evaluate IAF on three benchmarks with different modality hierarchies: text-dominant datasets (MultiHuSE, UR-FUNNY) and an audio-visual-dominant dataset (MUStARD). Pathway isolation shows that IAF preserves the dominant modality's internal accuracy at its unimodal ceiling across all tested configurations, whereas symmetric fusion degrades it by up to 18.5% on MultiHuSE. IAF improves over the strongest unimodal baseline by up to 8.25%.

## 1 Introduction

Early, late, and symmetric attention fusion form the standard toolkit for integrating text, audio, and video, resting on the intuition that combining heterogeneous signals should enable models to outperform any single modality alone (Baltrusaitis et al., 2019; Li and Tang, 2025). Yet on MultiHuSE (Kenneth et al., 2025b), we observe that these approaches often fail to surpass a text-only baseline and can even degrade performance in some settings. This pattern is not isolated: prior work in affective computing and action recognition reports similar findings, where multimodal models trained with conventional fusion strategies underperform their strongest unimodal component despite having access to additional modalities (Peng et al., 2022; Wang et al., 2020b).

Prior work attributes such failures to optimisation dynamics: modalities overfit and generalise at different rates, causing training to gravitate toward whichever modality converges fastest (Chaudhuri et al., 2025; Huang et al., 2022; Wang et al., 2020b). We extend this view by identifying an additional architectural mechanism that may contribute to the problem. Pathway isolation analysis of a symmetric attention model trained on MultiHuSE shows that the dominant text pathway drops from 74.9% to 56.4% after fusion. This substantial decline indicates that symmetric cross-attention forces the dominant modality to attend to weaker, noisier signals, degrading the representation that initially carried the most predictive information. We term this phenomenon *strong-modality collapse:* a failure mode in which the integration process itself undermines the model's reliance on its most informative modality.

This diagnosis suggests an architectural alternative. If symmetric attention contributes to this degradation, the solution lies in restructuring the fusion pathway rather than solely reweighting gradients or regularising losses. To this end, we propose Inverted Asymmetric Fusion (IAF), which imposes an explicit structural hierarchy. In IAF, the dominant modality bypasses cross-modal attention entirely to preserve its representation, while weaker modalities attend to it as a contextual anchor rather than acting as standalone classifiers. In addition, a modality-aware knowledge distillation stage (Hinton et al., 2015) precedes fusion, strengthening weaker modality representations so they can contribute a more informative complementary signal to the dominant pathway.

Testing across inverted modality hierarchies is essential to establish that IAF's structural approach generalises beyond any single dominance configuration. We therefore evaluate on MultiHuSE (Kenneth et al., 2025b) and UR-FUNNY (Kamrul Hasan et al., 2019), where text is dominant, and MUSTARD (Castro et al., 2019), where acoustic and visual cues carry more predictive signal. Our contributions are as follows:

- We identify strong-modality collapse as an architectural failure mode, distinct from optimisation-level modality imbalance, and provide pathway-level evidence of its occurrence in symmetric fusion.
- We propose IAF, a fusion framework that prevents strong-modality collapse by design and preserves the dominant modality's pathway accuracy at its unimodal ceiling across all tested configurations.
- We introduce a Modality-Aware Knowledge Distillation pipeline that adapts distillation to each dataset's modality hierarchy, strengthening weaker encoders prior to fusion.
- We demonstrate cross-hierarchy robustness on three benchmarks with differing dominance structures, improving over prior multimodal baselines on both UR-FUNNY and MUSTARD.

## 2 Related Work

### 2.1 Multimodal Fusion Architectures

Multimodal fusion has evolved from simple concatenation and prediction averaging to learned attention mechanisms (Baltrusaitis et al., 2019; Pu Liang et al., 2021). Tensor-based approaches explicitly model cross-modal interactions via outer products (Zadeh et al., 2017), with later work reducing computational cost through low-rank factorisation (Liu et al., 2018). The advent of attention-based fusion, led by the Multimodal Transformer (MulT) (Tsai et al., 2020) established symmetric cross-attention as the dominant paradigm. Subsequent improvements incorporated pretrained language models (Rahman et al., 2020), bottleneck token constraints (Nagrani et al., 2021), and modality subspace disentanglement (Hazarika et al., 2020).

While these designs capture rich cross-modal dynamics, they apply attention uniformly, regardless of the relative modality strengths (Feng et al., 2024). Forcing a dominant modality to symmetrically attend to weaker, noisier signals can corrupt its internal representation. This vulnerability motivates our *inverted asymmetric* approach, which preserves the dominant modality while using weaker streams as contextual anchors.

### 2.2 Modality Collapse and Pathway Analysis

Representational degradation in fusion is one instance of a broader failure mode: modality collapse. Empirical studies show multimodal models often exploit only a subset of their input modalities (Gong et al., 2025; Kenneth et al., 2024b; Zhang et al., 2024). Because different modalities overfit and generalise at uneven rates, training tends to favour the fastest-converging modality (Wang et al., 2020b; Wu et al., 2022). Theoretical analyses attribute this behaviour to conflicting gradient directions during joint optimisation (Huang et al., 2022; Chaudhuri et al., 2025).

To mitigate modality imbalance, prior work has explored optimisation-based strategies such as on-the-fly gradient modulation (OGM-GE) (Wei et al., 2025), Pareto-based integration (Wei and Hu, 2024), prototype-guided clustering (PMR) (Fan et al., 2023), and impartial multitask optimisation (Javaloy et al., 2022). These approaches primarily address the problem at the optimisation level by adjusting gradient signals or training objectives. However, they generally retain the same fusion topology, meaning cross-modal interactions can still expose dominant representations to interference. In contrast, our work explores a complementary architectural perspective. IAF restructures the fusion pathway to limit such interference by design. Nonetheless, structural shielding alone is insufficient unless weaker encoders are strengthened beforehand, motivating our pre-fusion distillation phase.

### 2.3 Knowledge Distillation in Multimodal Learning

Knowledge distillation (KD) transfers generalisation by aligning a student's predictions with a high-capacity teacher (Hinton et al., 2015; Gou et al., 2021). Cross-modal distillation extends this across domain boundaries and has proven effective for depth, optical flow, action recognition, and emotion analysis (Gupta et al., 2016; Thoker and Gall, 2019; Albanie et al., 2018). In multimodal fusion, KD is typically used for model compression (Fang et al., 2021; Wang et al., 2023) or robustness against

modality dropout (Wang et al., 2020a; Wei et al., 2023). More recent variants like DMD (Li et al., 2023b) transfer knowledge between shared and modality-exclusive subspaces, but still treat KD as a post-fusion refinement step. In contrast, we use KD pre-fusion to reduce modality gap before integration rather than relying on the architecture alone to compensate for encoder-level weaknesses.

### 2.4 Humor and Sarcasm Detection

The UR-FUNNY dataset (Kamrul Hasan et al., 2019) introduced large-scale multimodal humor detection, where textual signals dominate performance. Later work explored hierarchical punchline fusion (Choube and Soleymani, 2020) and optimal-transport-based alignment (Pramanick et al., 2022). MultiHuSE (Kenneth et al., 2025b) extends this to five-class humor style classification within a mental health context (Kenneth et al., 2024a, 2025a), again exhibiting a strong text bias. Such unbalanced modality hierarchies expose the limitations of symmetric fusion schemes that cannot maintain the integrity of the linguistic feature stream.

Conversely, MUStARD (Castro et al., 2019) reverses the hierarchy: sarcasm detection relies more on acoustic and visual cues, as textual sentiment is intentionally misleading. This structural inversion provides the ideal setting for evaluating IAF's cross-hierarchy robustness.

## 3 Methodology

Our framework comprises two sequential stages (Figure 1). **Stage 1** applies *Modality-Aware Knowledge Distillation* to strengthen weaker modalities prior to fusion, adapting the strategy to each dataset's modality hierarchy. **Stage 2** performs *Inverted Asymmetric Fusion* (IAF), integrating streams within a strict hierarchy that shields the dominant modality from cross-modal interference and prevents strong-modality collapse.

### 3.1 Problem Formulation

Let $\mathbf{f}^{(t)} \in \mathbb{R}^{d_t}$, $\mathbf{f}^{(a)} \in \mathbb{R}^{d_a}$, and $\mathbf{f}^{(v)} \in \mathbb{R}^{d_v}$ denote pre-extracted feature vectors for text, audio, and video modalities of a single utterance, respectively. Each is projected to a shared $d$=512 space via a modality-specific encoder $\phi_m$, yielding $\mathbf{h}^{(m)} = \phi_m(\mathbf{f}^{(m)}) \in \mathbb{R}^{512}$. The **dominant modality** $m^*$ is defined as $m^* = \arg\max_m$ Accuracy$(\phi_m)$ on the validation set.

### 3.2 Stage 1: Modality-Aware Knowledge Distillation

The dominant modality $m^*$ acts as a teacher to elevate weaker modalities $m \neq m^*$ before fusion.

#### 3.2.1 Distillation Loss ($\mathcal{L}_{\text{KD}}$).

Each student $\mathcal{S}$ is trained against a frozen teacher $\mathcal{T}$ (logits $\mathbf{z}^T \in \mathbb{R}^C$, bottleneck features $\mathbf{g}^T \in \mathbb{R}^{512}$) using:

$$\mathcal{L}_{\text{KD}} = \alpha\,\mathcal{L}_{\text{CE}} + \beta\,\mathcal{L}_{\text{soft}} + \gamma\,\mathcal{L}_{\text{feat}} \quad (1)$$

with fixed coefficients $\alpha = 0.4$, $\beta = 0.35$, $\gamma = 0.25$ that control the contribution of each component. These values were chosen to balance the three loss terms and kept constant across all datasets to evaluate generalisability.

The loss terms are: hard-label cross-entropy $\mathcal{L}_{\text{CE}}$, temperature-scaled KL divergence $\mathcal{L}_{\text{soft}}$ (with $T = 3.5$), and bottleneck feature alignment $\mathcal{L}_{\text{feat}}$:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} \mathbf{1}[y = c]\,\log\big(\text{softmax}(\mathbf{z}^S)_c\big) \quad (2)$$

$$\mathcal{L}_{\text{soft}} = T^2\,\text{KL}\left(\sigma\left(\frac{\mathbf{z}^S}{T}\right) \middle\| \sigma\left(\frac{\mathbf{z}^T}{T}\right)\right) \quad (3)$$

$$\mathcal{L}_{\text{feat}} = \left\|\mathbf{g}^S - \mathbf{g}^T\right\|_2^2 \quad (4)$$

where $\sigma$ denotes the softmax function.

**Weighted Dual-Teacher Distillation (MUStARD).** When both audio and video dominate (MUStARD), students are supervised by weighted dual teachers. The student is jointly supervised by a primary teacher $\mathcal{T}_1$ and a secondary teacher $\mathcal{T}_2$, whose losses are combined via a scalar weight $w_1$:

$$\mathcal{L}_{\text{dual}} = w_1\,\mathcal{L}_{\text{KD}}(\mathcal{T}_1) + (1 - w_1)\,\mathcal{L}_{\text{KD}}(\mathcal{T}_2) \quad (5)$$

where $w_1$ is tuned per configuration ($w_1 = 0.8$ for audio-primary, $w_1 = 0.2$ for video-primary).

**Cross-Architecture Distillation (UR-FUNNY).** On UR-FUNNY, the narrow unimodal performance gap limits the representational surplus available for cross-modal transfer. This motivated a same-modality, cross-architecture distillation: high-capacity static teachers (E5, Dasheng, PE-Core) distill into BiLSTM students on benchmark features (GloVe, COVAREP, OpenFace), isolating IAF's architectural contribution. The same three-term loss in Eq. (1) applies, with teacher features $\mathbf{g}^T$ projected to 512-dim prior to alignment.

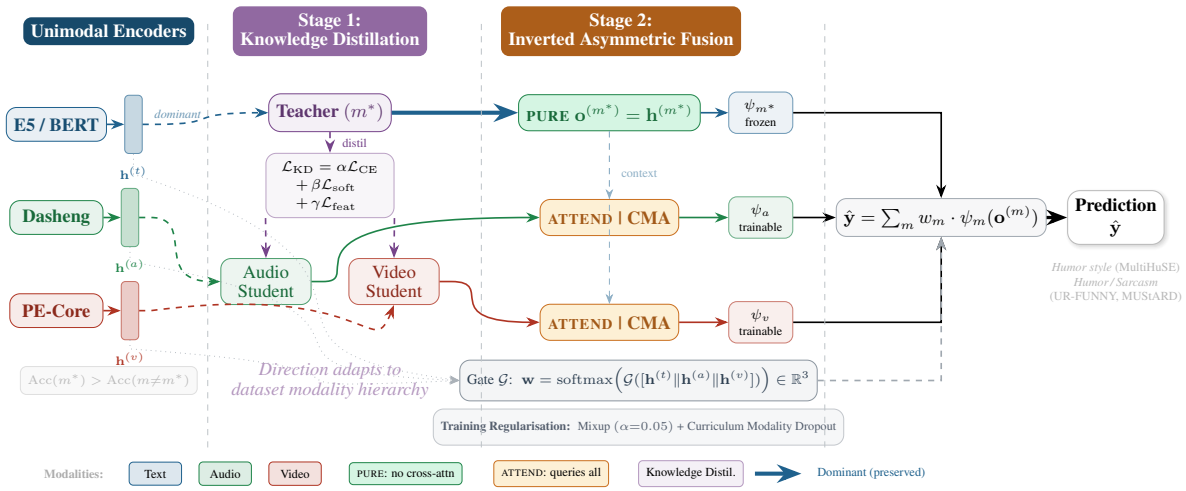


Figure 1: Overview of the proposed two-stage framework. Stage 1 performs Modality-Aware Knowledge Distillation, and Stage 2 applies the Inverted Asymmetric Fusion.

With all modality encoders operating at an enhanced representation, Stage 2 integrates these streams using IAF.

## 3.3 Stage 2: Inverted Asymmetric Fusion

### 3.3.1 CrossModal Anchor (CMA)

The *CrossModal Anchor* module serves as the core attention primitive in our architecture. Given a query-modality vector $\mathbf{h} \in \mathbb{R}^{512}$ and a context set $\mathbf{C} \in \mathbb{R}^{K \times 512}$ containing $K$ other modalities, the module computes:

$$\hat{\mathbf{h}} = \text{MHA}(\mathbf{h}, \mathbf{C}, \mathbf{C}) \tag{6}$$

$$\mathbf{h}' = \text{LayerNorm}\Big(\mathbf{h} + \text{Dropout}(\hat{\mathbf{h}})\Big) \tag{7}$$

where $\text{MHA}(\cdot)$ denotes multi-head attention (Vaswani et al., 2017) with $d_{\text{model}}$=512 and $H$=8 heads. The residual connection Eq. (7) preserves the query modality's original representation.

### 3.3.2 Structural Hierarchy: PURE vs. ATTEND

Let $m^*$ denote the dominant modality (or modalities). IAF enforces the following constraint on the forward pass:

$$\mathbf{o}^{(m)} = \begin{cases} \mathbf{h}^{(m)} & \text{if } m = m^* \text{ (PURE)} \\ \text{CMA}\Big(\mathbf{h}^{(m)},\ [\mathbf{h}^{(j)}]_{j \neq m}\Big) & \text{if } m \neq m^* \text{ (ATTEND)} \end{cases} \tag{8}$$

The PURE pathway remains unmodified. ATTEND pathways query all other modalities (including the shielded dominant) as contextual filters. On MultiHuSE/UR-FUNNY, text is PURE; on MUStARD, audio/video.

### 3.3.3 Adaptive Gating and Final Classification

Each modality produces pathway logits via a dedicated classifier head $\psi_m$. For the PURE modality, $\psi_{m^*}$ is a frozen copy of its standalone classifier, preserving its decision boundary. For ATTEND modalities, $\psi_m$ is a trainable linear head suited to their contextually-enriched representations.

A gating network $\mathcal{G}$ computes sample-adaptive weights from pre-attention features:

$$\mathbf{w} = \text{softmax}\Big(\mathcal{G}\Big([\mathbf{h}^{(t)} \,\|\, \mathbf{h}^{(a)} \,\|\, \mathbf{h}^{(v)}]\Big)\Big) \tag{9}$$

where $\mathbf{w} \in \mathbb{R}^3$ denotes the modality-gate weights, and $\mathcal{G}$ is a two-layer MLP with batch normalisation and dropout.

The final prediction is the weighted sum of per-pathway logits:

$$\hat{\mathbf{y}} = \sum_{m \in \{t,a,v\}} w_m \cdot \psi_m(\mathbf{o}^{(m)}) \tag{10}$$

Gating on pre-attention features preserves the PURE/ATTEND separation.

## 3.4 Training Regularisation

We apply two complementary regularisation strategies during fusion training.

### 3.4.1 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:

$$\tilde{\mathbf{h}}_i^{(m)} = \lambda\, \mathbf{h}_i^{(m)} + (1-\lambda)\, \mathbf{h}_{\sigma(i)}^{(m)}, \quad \tilde{y}_i = (\lambda,\ y_i,\ y_{\sigma(i)}) \tag{11}$$

where $\sigma$ is a random permutation of batch indices. Mixup discourages overconfident predictions, which is critical for preventing overfitting (ablation in Appendix C)

#### 3.4.2 Curriculum Modality Dropout

To prevent over-reliance on any single modality, we apply a two-phase stochastic masking schedule: all modalities are present during warm-up (epochs <15), after which each sample is assigned one of seven masking configurations (trimodal $p = 0.40$, each bimodal pair $p = 0.10$, each unimodal $p = 0.10$), with masked modalities zeroed before the forward pass.

### 3.5 Pathway Analysis

To verify dominant pathway preservation, we isolate each modality's internal contribution by routing $\mathbf{h}^{(m)}$ directly through $\psi_m$, bypassing gating.

For the PURE modality $m^*$:

$$\text{Accuracy}_{\text{path}}^{(m^*)} = \text{Accuracy}_{\text{uni}}^{(m^*)} \tag{12}$$

That is, $m^*$ pathway accuracy equals its standalone unimodal performance. Any deviation from this equality indicates cross-modal interference. Symmetric fusion violates this equality; IAF enforces it by construction. For ATTEND modalities, pathway accuracy is expected to drop below their unimodal baselines, as their representations are no longer optimised for standalone classification but instead tuned to provide contextual signal.

## 4 Experiments

We describe the datasets, unimodal feature extraction procedures, baseline models, and implementation details.

### 4.1 Datasets

We evaluate on three multimodal benchmarks for humor and sarcasm detection, chosen to span contrasting modality hierarchies and task complexity (binary and multiclass classification).

**MultiHuSE** (Kenneth et al., 2025b) is a 5-class humor style classification dataset comprising 2,407 video utterances from 50 diverse actors, including 1,463 unique text scripts and 943 re-performances. Standard random splits risk data leakage across performances of the same script, so we introduce a strict 5-fold cross-validation protocol grouped by unique text ID, ensuring all performances of a given script appear exclusively in either training or evaluation. All baseline models are re-implemented and re-evaluated under this protocol. Modality hierarchy: Text $\gg$ Audio $\gg$ Video.

**UR-FUNNY** (Kamrul Hasan et al., 2019) is a binary humor detection dataset of 16,514 context–punchline pairs drawn from TED talks, using the standard train/dev/test split ($N_{\text{test}}$=994). Following the original benchmark, we use sequential features GloVe (text), COVAREP (audio), and OpenFace (video) throughout, isolating our architectural contribution from feature quality and enabling direct comparison with prior work using the same feature set. Modality hierarchy: Text > Audio $\approx$ Video.

**MUStARD** (Castro et al., 2019) is a binary sarcasm detection dataset of 690 balanced utterances drawn from TV sitcoms, using 5-fold cross-validation ($N_{\text{test}}$=138 per fold) from the original paper. Modality hierarchy: Audio $\approx$ Video > Text, the most strongly inverted relative to MultiHuSE, and the primary motivation for our cross-hierarchy evaluation.

### 4.2 Feature Extraction

All features are extracted offline using frozen pretrained encoders and serialised to disk prior to training. This design decouples representation quality from the fusion architecture, ensuring that observed performance differences reflect architectural choices rather than effects of encoder fine-tuning.

**MultiHuSE and MUStARD.** For text, we evaluate two encoders: BERT-based-uncased (Devlin et al., 2019) (768-d) and Multilingual-E5-large-instruct (Wang et al., 2024) (1024-d). Using both allows us to verify that our findings are not specific to a single text backbone. For audio, we use Dasheng-0.6B (Dinkel et al., 2024) (1280-d), a large-scale general-purpose audio encoder. For video, we evaluate two encoders: VideoMAE (Tong et al., 2022) (768-d) and PE-Core (Bolya et al., 2025) (1024-d), providing coverage of both masked autoencoder and contrastive video representations.

**UR-FUNNY.** We use GloVe (Pennington et al., 2014) (text), COVAREP (Degottex et al., 2014) (audio), and OpenFace (Baltrusaitis et al., 2018) (video), consistent with the original benchmark. The role of modern encoders as distillation teachers within this established feature setting is described in Section 3.2.

### 4.3 Implementation Details

**Infrastructure.** Feature extraction is performed on an HPC cluster (NVIDIA A16, 15GB VRAM); all subsequent training uses a consumer laptop (NVIDIA GeForce RTX 3050, 4GB VRAM). All experiments use a fixed random seed of 999.

**Model Parameters.** Pre-trained encoders used for feature extraction have the following parameter counts: BERT-base (110M), E5-multilingual (560M), Dasheng (600M), VideoMAE-base (94M), and PE-Core (320M visual encoder + 310M language tower). These encoders are frozen throughout; no encoder parameters are updated during distillation or fusion training.

**Distillation and Fusion.** All models use AdamW for distillation, CosineAnnealingLR for fusion, a batch size of 64, Mixup with $\alpha_{\text{mix}} = 0.05$, a curriculum dropout warm-up of 15 epochs, and a CrossModal Anchor with $d_{\text{model}} = 512$ and $H = 8$ heads. Full hyperparameters are provided in Appendix A.

**Evaluation.** We report accuracy and macro-averaged F1. MultiHuSE and MUStARD scores are mean±SD over 5 folds; UR-FUNNY follows the standard single-split protocol.

### 4.4 Baselines

We compare against three internal baselines trained on identical features under identical hyperparameters: (1) **Early Fusion**: feature concatenation followed by an MLP; (2) **Late Fusion**: weighted averaging of unimodal prediction vectors; (3) **Symmetric Fusion**: mutual cross-attention across all three modalities using the same distilled features as IAF, isolating the inverted asymmetric design. Comparisons with prior published work appear in the results tables in Section 5.

## 5 Results and Analysis

Table 1 reports unimodal performance. Distillation consistently improves student models (gains of 0.88–6.23%, mean 2.44%), with the largest gain on UR-FUNNY's COVAREP audio (+6.23%), highlighting the benefit of cross-architecture distillation. These distilled students, together with their teachers, serve as the inputs to IAF and symmetry fusion.

Table 2 presents trimodal fusion results on MultiHuSE. Early and late fusion fail to reliably surpass the dominant unimodal text baseline, with early fusion degrading sharply in the E5+VideoMAE setting (−10.55%), suggesting that naive concatenation can disrupt strong text representations when combined with noisier video inputs. Symmetric fusion also fails to recover the E5 baseline in both E5 configurations (−2.49%, −0.54%), indicating that expressive attention alone does not protect the dominant modality. IAF is the only architecture that consistently exceeds the unimodal ceiling across all four configurations, attaining gains of 2.06–6.48%.

| Modality / Feature | Base Acc(%) | Base F1(%) | Dist. Acc(%) | Dist. F1(%) |
|---|---|---|---|---|
| ***MultiHuSE*** *— 5-class humor style \| Text ≫ Audio ≫ Video* | | | | |
| 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 |
| ***UR-FUNNY*** *— binary humor \| Text > Audio ≈ Video* | | | | |
| Text: GloVe* | 62.47 | 64 | **65.19** | **65** |
| Audio: COVAREP | 57.55 | 48 | 63.78 | 64 |
| Video: OpenFace | 57.75 | 58 | 58.95 | 59 |
| ***MUStARD*** *— binary sarcasm \| Audio ≈ Video > Text* | | | | |
| 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 | – | – |

Table 1: Unimodal results across all three benchmarks. *Base*: raw pre-trained; *Dist.*: knowledge-distilled. * = dominant modality. MultiHuSE and MUStARD: mean ± SD over 5-fold CV; UR-FUNNY: single split.

| Text | Video | Architecture | Acc | F1 | Δ |
|---|---|---|---|---|---|
| *Audio: Dasheng used across all configurations* | | | | | |
| *Text: BERT (base 68.26%)* | | | | | |
| BERT | VideoMAE | Early Fusion | 67.22 ±2.00 | 67 | −1.04 |
| | | Late Fusion | 68.05 ±2.12 | 68 | −0.21 |
| | | Symmetric Fusion | 69.42 ±1.48 | 69 | +1.16 |
| | | **Proposed (IAF)** | **73.16 ±2.59** | **72** | +4.90 |
| BERT | PE-Core | Early Fusion | 70.38 ±2.57 | 70 | +2.12 |
| | | Late Fusion | 68.63 ±2.36 | 68 | +0.37 |
| | | Symmetric Fusion | 72.00 ±1.20 | 70 | +3.74 |
| | | **Proposed (IAF)** | **74.74 ±1.71** | **74** | +6.48 |
| *Text: E5 (base 74.86%)* | | | | | |
| E5 | VideoMAE | Early Fusion | 64.31 ±1.51 | 64 | −10.55 |
| | | Late Fusion | 71.25 ±4.07 | 71 | −3.61 |
| | | Symmetric Fusion | 72.37 ±2.73 | 72 | −2.49 |
| | | **Proposed (IAF)** | **76.92 ±3.53** | **76** | +2.06 |
| E5 | PE-Core | Early Fusion | 76.28 ±2.91 | 76 | +1.42 |
| | | Late Fusion | 71.54 ±4.20 | 71 | −3.32 |
| | | Symmetric Fusion | 74.32 ±3.03 | 74 | −0.54 |
| | | **Proposed (IAF)** | **78.06 ±3.82** | **78** | +3.20 |

Table 2: Trimodal fusion results on MultiHuSE. Δ = gain over the strongest unimodal text baseline (BERT = 68.26%; E5 = 74.86%). All metrics in %.

Table 3 compares IAF against prior work and our internal baselines on UR-FUNNY. IAF achieves 70.72% accuracy, surpassing the strongest prior

method (MISA, 68.60%) by 2.12% and symmetric fusion by 1.81%. Interestingly, symmetric fusion (68.91%) already outperforms previously published methods, showing that strengthening weaker modalities before fusion yields substantial gains even without architectural asymmetry.

| Method | Context | Acc (%) | F1 (%) | $\Delta_{\text{uni}}$ |
|---|---|---|---|---|
| *Prior published methods* | | | | |
| 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 and Soleymani, 2020) | ✓ | 67.84 | 69 | – |
| MISA (GloVe) (Hazarika et al., 2020) | × | 68.60 | – | – |
| *Our internal baselines — Text: GloVe (base 62.47%)* | | | | |
| Early Fusion | ✓ | 65.19 | 66 | +2.72 |
| Late Fusion | ✓ | 64.19 | 65 | +1.72 |
| Symmetric Fusion | ✓ | 68.91 | 69 | +6.44 |
| **Proposed (IAF)** | ✓ | **70.72** | 71 | +8.25 |

Table 3: Results on UR-FUNNY. All methods use GloVe/COVAREP/OpenFace features (Kamrul Hasan et al., 2019). $\Delta_{\text{uni}}$ = gain over the GloVe baseline (62.47%) for our internal baselines. Context: ✓ = punchline + context; × = punchline only.

Table 4 reports results on MUStARD. Here, symmetric fusion already improves markedly over the dominant unimodal audio baseline (+3.47% and +4.34% for BERT and E5, respectively), yet IAF further boosts performance in both text-encoder settings. In particular, IAF (E5) attains 83.33% accuracy, exceeding the state-of-the-art method (MHA with RoBERTa, 79.32%) by 4.01%.

| Method | Context | Acc(%) | F1(%) | $\Delta_{\text{uni}}$ |
|---|---|---|---|---|
| *Prior published methods* | | | | |
| 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 | – |
| *Our internal baselines – Audio: Dasheng (base 78.12%)* | | | | |
| Early Fusion | ✓ | 79.71 ±3.04 | 80.0 | +1.59 |
| Late Fusion (Average) | ✓ | 79.86 ±4.71 | 80.0 | +1.74 |
| Symmetric Fusion (BERT) | ✓ | 81.59 ±4.17 | 82.0 | +3.47 |
| Symmetric Fusion (E5) | ✓ | 82.46 ±4.52 | 82.0 | +4.34 |
| Proposed (IAF (BERT)) | ✓ | 82.61 ±3.92 | 83.0 | +4.49 |
| **Proposed (IAF (E5))** | ✓ | **83.33 ±4.20** | **84.0** | +5.21 |

Table 4: Results on MUStARD. $\Delta_{\text{uni}}$ = gain over Dasheng audio unimodal baseline (our internal baselines only).

Table 5 summarises pathway isolation analysis. Across all three datasets, IAF preserves the dominant modality's pathway at its unimodal baseline, whereas symmetric fusion degrades the same pathways by up to 18.5% (E5 on MultiHuSE: 74.9% → 56.4%). This directly supports our claim that IAF's structural asymmetry functions as a representation-preservation guarantee that symmetric fusion cannot provide.

| Text | Video | Archi. | Full Fusion | Text Path | Audio Path | Video Path |
|---|---|---|---|---|---|---|
| *MultiHuSE – Dominant Modality: TEXT – BERT=70.2%, E5=74.9%* | | | | | | |
| BERT | VideoMAE | Symmetric | 69.42±1.48 | 61.3±2.5 | 59.0±4.5 | 60.1±1.3 |
| | | IAF | 73.16±2.59 | **70.2±2.8** | 55.8±7.4 | 57.5±5.0 |
| | PE-Core | Symmetric | 72.00±1.20 | 63.5±1.6 | 63.6±2.3 | 61.6±7.2 |
| | | IAF | 74.74±1.71 | **70.2±2.8** | 62.9±2.5 | 65.6±3.1 |
| E5 | VideoMAE | Symmetric | 72.37±2.73 | 56.4±4.6 | 63.7±9.2 | 67.4±3.3 |
| | | IAF | 76.92±3.53 | **74.9±3.1** | 48.4±8.9 | 56.7±7.0 |
| | PE-Core | Symmetric | 74.32±3.03 | 55.3±8.7 | 65.5±2.2 | 69.7±4.8 |
| | | IAF | 78.06±3.82 | **74.9±3.1** | 54.9±9.6 | 62.2±6.6 |
| *UR-FUNNY – Dominant Modality: TEXT=65.19%* | | | | | | |
| GloVe | OpenFace | Symmetric | 68.91 | 63.08 | 49.30 | 50.70 |
| | | IAF | 70.72 | **65.19** | 50.91 | 49.30 |
| *MUStARD– Dominant Modalities: AUDIO=78.12%; VIDEO (PE)=77.10%* | | | | | | |
| BERT | VideoMAE | Symmetric | 79.28±4.26 | 62.0±5.0 | 64.4±15.2 | 67.8±8.0 |
| | | IAF | 80.43±2.63 | 64.6±9.0 | **78.1±5.5** | **74.5±1.2** |
| | PE-Core | Symmetric | 81.59±4.17 | 64.9±9.5 | 69.4±6.3 | 73.5±9.0 |
| | | IAF | 82.61±3.92 | 63.9±6.2 | **78.1±5.5** | **77.1±3.1** |
| E5 | VideoMAE | Symmetric | 79.86±3.09 | 60.3±12.3 | 77.3±5.1 | 71.5±6.2 |
| | | IAF | 79.42±4.34 | 66.4±7.6 | **78.1±5.5** | **74.5±1.2** |
| | PE-Core | Symmetric | 82.46±4.52 | 74.4±3.5 | 74.9±4.6 | 73.3±10.4 |
| | | IAF | 83.33±4.20 | 69.0±7.8 | **78.1±5.5** | **77.1±3.1** |

Table 5: Pathway isolation analysis. Each score reflects accuracy using only the corresponding modality pathway, bypassing fusion aggregation. **Bold** = dominant modality pathway. All metrics in %.

## 5.1 Ablation Study

**(1) Role of Modalities.** We remove one modality at a time to quantify each modality's contribution to full trimodal performance. Table 6 shows that full trimodal IAF outperforms all bimodal variants on every dataset. On MultiHuSE, the text-free A+V pair drops to 65.85%, confirming text's dominant role. On UR-FUNNY, $T{+}A$ outperforms $T{+}V$ (68.91% vs. 65.29%), consistent with the Text > Audio ≈ Video hierarchy. On MUStARD, the text-free A+V pair (82.61%) nearly matches full trimodal performance (83.33%), reflecting the co-dominance of audio and video and the smaller marginal benefit of text.

**(2) Role of Distilled Feature Representations.** We contrast IAF trained on raw pre-trained features with IAF trained on distillation-enhanced features, isolating the impact of Stage 1. As shown in Table 7, incorporating distilled representations consistently improves performance: gains of 1.02–2.41% on MultiHuSE, 1.02–1.59% on MUStARD, and 1.81% on UR-FUNNY. This underscores the importance of raising weaker modality representations

| 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 | **70.72** | **71** |
| 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 6: Bimodal ablation of IAF (best encoder per dataset). *Pure* = dominant modality; *Attends* = contextual filters. BERT-based configurations are reported in Appendix B, Table 10.

before fusion, rather than relying solely on the architecture.

| Dataset | Config | Feat. | Acc(%) ± SD | F1(%) | Δ(%) |
|---|---|---|---|---|---|
| MultiHuSE | BERT+PEC | Base | 72.33±2.02 | 72 | -2.41 |
| | | Dist | **74.74**±1.71 | **74** | – |
| | E5+PEC | Base | 75.90±6.17 | 76 | -2.16 |
| | | Dist | **78.06**±3.82 | **78** | – |
| MUStARD | BERT+PEC | Base | 81.59±3.85 | 82 | -1.02 |
| | | Dist | **82.61**±3.92 | **83** | – |
| | E5+PEC | Base | 81.74±4.24 | 82 | -1.59 |
| | | Dist | **83.33**±4.20 | **84** | – |
| UR-FUNNY | GloVe/COV/OF | Base | 68.91 | 69 | -1.81 |
| | | Dist | **70.72** | **71** | – |

Table 7: Distillation ablation: raw (Base) vs. distilled (Dist.) features within IAF. $\Delta$ = Dist − Base.

**(3) Role of Regularisation.** Appendix C, Table 11, reports the contribution of Mixup and Curriculum Modality Dropout. Removing both yields the largest drops across all datasets (up to 4.07% on MultiHuSE, 1.61% on UR-FUNNY, and 2.90% on MUStARD), confirming their complementary roles. Curriculum Dropout provides a consistently larger isolated gain than Mixup, suggesting that progressive difficulty scheduling is particularly effective in noisy multimodal feature spaces.

**(4) Role of the Gate Mechanism.** We evaluate three gating strategies within IAF: a *Learnable* gate (sample-adaptive weights), a *Hierarchy* gate (fixed weights derived from modality ranking), and a *Uniform* gate (equal weights, T=A=V=0.33) Results (Appendix D, Table 12) show that the learnable gate usually performs best, while fixed hierarchical weights are competitive under extreme modality imbalance, indicating that IAF remains robust to the specific gating choice.

## 6 Conclusion

In this work, we introduced Inverted Asymmetric Fusion (IAF), a two-stage multimodal framework designed to address strong-modality collapse in discriminative fusion models. By structurally enforcing a PURE/ATTEND hierarchy, IAF shields the dominant modality from cross-modal interference while allowing weaker modalities to act as contextual filters. A Modality-Aware Knowledge Distillation pipeline, adapted per dataset, further ensures that each modality stream enters fusion at its representational ceiling. Post-hoc pathway isolation analysis confirms that this structural design preserves the dominant unimodal baseline, which standard symmetric cross-attention tends to degrade.

Evaluated across three benchmarks spanning different modality dominance structures: text-dominant MultiHuSE and UR-FUNNY, and audio–visual-dominant MUStARD, IAF achieves consistent gains over early, late, and symmetric fusion baselines, and surpasses prior published methods on both UR-FUNNY and MUStARD. The cross-hierarchy consistency of these results suggests that modality preservation, rather than increased fusion expressivity alone, is a key factor in robust multimodal learning. We hope this work encourages greater attention to architectural asymmetry as a principled response to modality imbalance in multimodal systems, and to the careful design of fusion schemes for deployed multimodal models.

## Limitations

Several limitations of the present work should be noted. First, IAF requires the dominant modality to be identified empirically via unimodal evaluation prior to training. In settings where modalities possess comparable predictive strength, the PURE/ATTEND assignment becomes ambiguous and the benefits of structural asymmetry may be reduced. While the dual-teacher scheme used for MUStARD partially addresses co-dominant settings, a more principled treatment of this edge case remains an open problem.

Second, all encoders are used in a frozen, pre-extracted configuration. This choice deliberately isolates the architectural contribution of IAF from encoder fine-tuning effects, but the framework has not been evaluated in an end-to-end setting, where

joint fine-tuning could alter the modality hierarchy and the effectiveness of the distillation strategy.

Finally, all datasets considered are English-language only. The inclusion of E5-multilingual in several configurations suggests potential cross-lingual applicability, but this has not been empirically tested. The extent to which our modality hierarchy observations and IAF's benefits transfer to non-English or low-resource multimodal settings is an important direction for future work.

## Ethical considerations

This work uses three publicly available datasets (MultiHuSE, UR-FUNNY, MUStARD), collected and released under standard academic licences by their respective authors. No new data collection or human annotation was conducted. All experiments operate on pre-extracted, anonymised feature representations with no access to personally identifiable information. We note that automated humor and sarcasm detection systems may misclassify culturally specific communication styles and are not intended for deployment beyond controlled research settings.

## References


Sajal Aggarwal, Ananya Pandey, and Dinesh Kumar Vishwakarma. 2023. 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.

Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, pages 292–301. Association for Computing Machinery, Inc.

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

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

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. 2025. Perception Encoder: The best visual embeddings are not at the output of the network. *arXiv Preprint (Facebook research)*.

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

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. 2025. A Closer Look at Multimodal Representation Collapse. In *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver.

Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. 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*, pages 4351–4360. Association for Computational Linguistics.

Akshat Choube and Mohammad Soleymani. 2020. Punchline Detection using Context-Aware Hierarchical Multimodal Fusion. In *ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 675–679. Association for Computing Machinery, Inc.

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

Boya Deng, Jianzhao Li, Maoguo Gong, Zedong Tang, Yourun Zhang, Kaiyuan Feng, and Yue Wu. 2025. BCFNet: Bi-temporal collaborative fusion network for multi-modal humor detection. *Pattern Recognition*, 172.

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

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

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. PMR: Prototypical Modal Rebalance for Multimodal Learning. In

*IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20029–20038, Vancouver. IEEE.

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

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

Baoquan Gong, Xiyuan Gao, Pengfei Zhu, Qinghua Hu, and Bing Cao. 2025. Multimodal Negative Learning. In *39th Conference on Neural Information Processing Systems*. NeurIPS.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6):1789–1819.

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

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv Preprint*.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. 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.

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

Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Hoque. 2019. 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*, pages 2046–2056.

Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. 2024a. A Two-Model Approach for Humour Style Recognition. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 259–274, Miami. Association for Computational Linguistics.

Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. 2024b. Systematic Literature Review: Computational Approaches for Humour Style Classification. *ArXiv*.

Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. 2025a. Explaining Humour Style Classifications: An XAI Approach to Understanding Computational Humour Analysis. *Journal of Data Mining & Digital Humanities, NLP4DH*.

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

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

Songtao Li and Hao Tang. 2025. Multimodal Alignment and Fusion: A Survey. *International Journal of Computer Vision*.

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

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

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

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8238–8247.

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

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

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. 2021. MULTIBENCH: Multiscale Benchmarks for Multimodal Representation Learning. In *35th Conference on Neural Information Processing Systems, Track on Datasets and Benchmarks.*, pages 1–20. NeurIPS.

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

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. 2022. A Multimodal Corpus for Emotion Recognition in Sarcasm. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 20–25.

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

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. 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.

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

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint*.

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

Tiannan Wang, Wangchunshu Zhou, Eth Zurich, Yan Zeng, and Xinsong Zhang. 2023. EfficientVLM: Fast and Accurate Vision-Language Models via Knowledge Distillation and Modal-adaptive Pruning. In *Findings of the Association for Computational Linguistics*, pages 13899–13913. Association for Computational Linguistics.

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

Shicai Wei, Chunbo Luo, and Yang Luo. 2023. 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)*, pages 20039–20049. IEEE.

Yake Wei and Di Hu. 2024. MMPareto: Boosting Multimodal Learning with Innocent Unimodal Assistance. In *Proceedings of the 41st International Conference on Machine Learning*, Vienna.

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

Nan Wu, Stanisław Jastrz̨ Ebski, Kyunghyun Cho, and Krzysztof J Geras. 2022. Characterizing and Overcoming the Greedy Nature of Learning in Multimodal Deep Neural Networks. In *Proceedings of the 39th International Conference on Machine Learning*.

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

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

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024. Multimodal Representation Learning by Alternating Unimodal Adaptation. In *IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*. Computer Vision and Pattern Recognition Conference (CVPR).

## A Full Hyperparameter Configurations

Table 9 and Table 8 present the complete hyperparameter configurations for the distillation and fusion stages, respectively. Hyperparameters were selected through manual tuning: values were adjusted iteratively based on validation performance until stable, well-performing configurations were identified, after which a single final experiment was run per configuration and reported in the paper. No automated search was conducted.

| Hyperparameter | MultiHuSE | UR-FUNNY | MUStARD |
|---|---|---|---|
| Optimiser | AdamW | Adam | AdamW |
| Learning rate | $6 \times 10^{-5}$ | $1 \times 10^{-5}$ | $1 \times 10^{-3}$ |
| Weight decay | $10^{-3}$ | $10^{-3}$ | $10^{-2}$ |
| Max epochs | 100 | 80 | 100 |
| Early stopping patience | 35 | 20 | 40 |
| Gradient clipping | 1.0 | — | 1.0 |

Table 8: Dataset-specific hyperparameters for the fusion

| Hyperparameter | Value |
|---|---|
| KD loss weights $(\alpha, \beta, \gamma)$ | 0.4, 0.35, 0.25 |
| Temperature $T$ | 3.5 |
| Optimiser | AdamW |
| Weight decay | $10^{-3}$ |
| Text/Audio LR | $10^{-4}$–$10^{-3}$ |
| Video LR | $9 \times 10^{-5}$ |

Table 9: Distillation stage hyperparameters (shared across all datasets).

## B Bimodal Ablation: BERT

Table 10 reports bimodal ablation results for BERT encoder configurations on MultiHuSE and MUStARD. Results follow the same directional patterns as the E5 configurations in the main paper (Table 6), confirming that the modality hierarchy findings are consistent across text encoders.

| 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 |

Table 10: Bimodal fusion for BERT configurations on MultiHuSE and MUStARD. Directional patterns are consistent with E5 configurations in Table 6.

## C Role of Regularisation

Table 11 reports the full regularisation ablation across all datasets and encoder configurations.

| 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** *(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 11: Regularisation 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 %.

BERT- and E5-based configurations are shown for MultiHuSE and MUStARD; results are directionally consistent across both encoders.

## D Role of Gate Mechanism

Table 12 reports the effect of different gating strategies. The learnable gate attains the highest accuracy on UR-FUNNY and MUStARD, demonstrating the benefit of sample-adaptive weighting. MultiHuSE is a mild exception where the fixed hierarchy gate performs marginally better, likely due to its extreme modality imbalance: under highly skewed conditions, a learnable gate can over-emphasize the dominant modality. Overall differences between gates are small ($\leq$1.45%), indicating that IAF is largely insensitive to the precise gating strategy.

| 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* (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 |

Table 12: Gate mechanism ablation across all 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.