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).

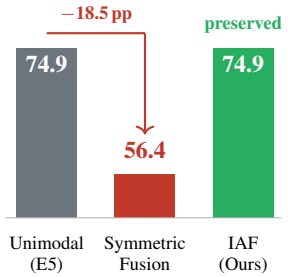

Figure 1: Text-pathway accuracy on MultiHuSE (E5 encoder). Symmetric fusion degrades the dominant modality by 18.5 pp; IAF preserves it (Table 5).

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 on MultiHuSE shows that the dominant text pathway drops from 74.9% to 56.4% after fusion (Figure 1).

This decline suggests 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 fusion undermines the model's strongest 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 gra-

dients or losses (Wei et al., 2025; Fan et al., 2023; Javaloy et al., 2022), which adjust training dynamics while leaving the cross-attention topology intact. 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 as standalone classifiers. Additionally, 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-based methods that reweight gradients while preserving symmetric cross-attention, and provide pathway-level evidence 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 across 3 benchmarks with differing dominance structures, improving on prior baselines on 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. Recent work has begun to address this by using text to guide multimodal fusion in affective tasks (Zhang et al., 2025), though such designs typically retain mutual attention rather than fully shielding the dominant stream from cross-modal interference. 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 proposed 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 methods address imbalance by adjusting gradients or training objectives, but generally retain the same fusion topology, leaving dominant representations vulnerable to cross-modal interference. In contrast, our work takes a complementary architectural approach. IAF restructures the fusion pathway to limit such inter-

ference by design. However, structural shielding alone is insufficient unless weaker encoders are first strengthened, motivating the pre-fusion distillation stage.

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 commonly used within the fusion stage for compression, dropout robustness, missing-modality learning, or representational alignment (Fang et al., 2021; Wang et al., 2023, 2020a; Wei et al., 2023; Li et al., 2023b; Lin and Hu, 2024). By contrast, our Modality-Aware Knowledge Distillation (MAKD) is a pre-fusion stage that uses only the empirically dominant unimodal modality as teacher and is designed specifically to strengthen weaker encoders before asymmetric integration. A detailed design comparison is provided in Appendix H.

2.4 Humor and Sarcasm Detection

Humor and sarcasm benchmarks provide a useful testbed for modality imbalance because their predictive hierarchies differ substantially across datasets. UR-FUNNY (Kamrul Hasan et al., 2019) and MultiHuSE (Kenneth et al., 2025b, 2024a, 2025a) are text-dominant, whereas MUStARD (Castro et al., 2019) places greater weight on acoustic and visual cues. This contrast makes the domain well suited for evaluating whether a fusion architecture can preserve strong modalities without sacrificing cross-modal contribution.

3 Methodology

Our framework comprises two sequential stages (Figure 2). **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 ϕ_m, yielding $\mathbf{h}^{(m)} = \phi_m(\mathbf{f}^{(m)}) \in \mathbb{R}^{512}$. The **dominant modality** m^* is defined as $m^* = \arg\max_m \text{ Accuracy}(\phi_m)$ 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:

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

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 $\mathcal{L}_{\text{feat}}$:

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

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

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

where σ denotes the softmax function. We use fixed coefficients $\alpha = 0.4$, $\beta = 0.35$, and $\gamma = 0.25$, selected by manual tuning on MultiHuSE and held constant across all datasets and encoder combinations. A sensitivity sweep of coefficients over four configurations shows that fusion accuracy varies by at most 1.20 percentage points, with both the default and equal-weight settings achieving the best result; see Appendix G for details.

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:

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

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

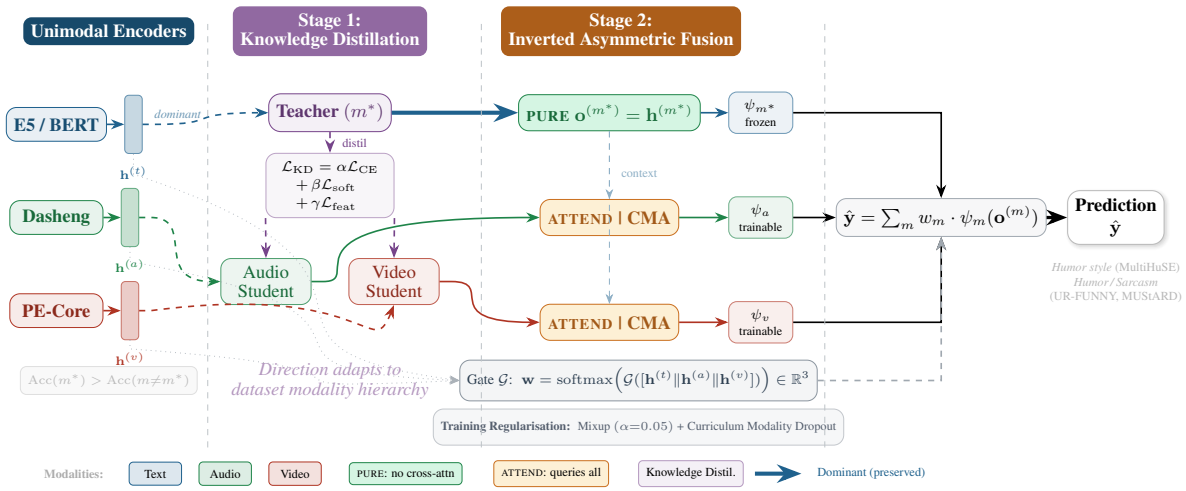

Figure 2: Overview of the proposed two-stage framework. Stage 1 performs Modality-Aware Knowledge Distillation, and Stage 2 applies the Inverted Asymmetric Fusion.

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.

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_{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 ψ_m. For the PURE modality, ψ_{m^*} is a frozen copy of its standalone classifier, preserving its decision boundary. For ATTEND modalities, ψ_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 textscpure/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 α_{mix}=0.05 and interpolate pairs of feature vectors and their labels:

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

where σ 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 ψ_m, bypassing gating. For the PURE modality m^*:

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

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, 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_{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_{test}=138 per fold) from the original paper. Modality hierarchy: Audio $\approx$ Video > Text, is the clearest inversion relative to MultiHuSE and motivates 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), while all subsequent training is conducted on a consumer laptop (NVIDIA GeForce RTX 3050, 4GB VRAM). All primary results use a fixed random seed of 999. For MultiHuSE and MUStARD, results are reported as mean $\pm$ SD over 5-fold cross-validation; for UR-FUNNY, we additionally report a five-seed analysis using seeds 42, 100, 123, 999, and 600. The unimodal pre-evaluation used to identify the dominant modality adds minimal overhead: on our hardware, training a single unimodal encoder takes less than 15 minutes, compared with a few hours for full fusion training.

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. We used 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. For MultiHuSE and MUStARD, results are reported as mean $\pm$ SD over 5 folds. For UR-FUNNY, we use the standard train/dev/test split; the result at seed 999 is reported in Table 3 for direct comparability with prior published baselines, and a multi-seed analysis is reported in Section 5.

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.

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 $\pm$ SD over 5-fold CV; UR-FUNNY: single split.

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 may 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 is insufficient to preserve 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%. Paired Wilcoxon signed-rank tests across the five folds show that IAF significantly outperforms symmetric fusion in the PE-Core configurations (BERT: $W{=}0$, $p{=}0.031$; E5: $W{=}0$, $p{=}0.031$; one-tailed), a result further supported by paired t-tests ($p{=}0.023$ and $p{=}0.005$, respectively); Here, $W{=}0$ indicates IAF outperformed symmetric fusion on every fold.

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 %.

Method	Context	Acc (%)	F1 (%)	Δ_{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). Δ_{uni} = gain over the GloVe baseline (62.47%) for our internal baselines. Context: ✓ = punchline + context; × = punchline only. †Seed 999; mean 69.70 ± 0.61% over 5 seeds.

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%) outperforms previously published methods, demonstrating that strengthening weaker modalities before fusion is beneficial even without architectural asymmetry. To assess initialisation sensitivity, we evaluated IAF over five random seeds (42, 100, 123, 600, 999), obtaining a mean accuracy of 69.70 ± 0.61% (min 69.32%; seed 999 is reported in Table 3 for comparability with prior work). A one-sample t-test shows that this mean significantly exceeds both symmetric fusion ($t(4)$=2.90, p=0.022, one-tailed) and MISA ($t(4)$=4.03, p=0.008, one-tailed), and each seed individually outperforms both baselines.

Table 4 reports results on MUStARD. Symmetric fusion already improves substantially over the strongest unimodal audio baseline (+3.47% and +4.34% for BERT and E5, respectively), and IAF yields further gains 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%. We note that the IAF-over-symmetric margin on MUStARD (0.87–1.02 pp) falls within one standard deviation of both models and should be interpreted accordingly. Paired Wilcoxon signed-rank tests are consistent with this pattern: neither BERT+PE-Core (W=2, p=0.188) nor E5+PE-Core (W=1, p=0.125) is significant (one-tailed, α=0.05).

Method	Context	Acc(%)	F1(%)	Δ_{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. Δ_{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 18.5pp: the E5 text pathway drops from 74.9% (Table 1) to 56.4% under symmetric fusion (Table 5). This directly supports our claim that IAF's structural asymmetry functions as a representation-preservation guarantee that symmetric fusion cannot provide. Visual summaries of this collapse pattern are provided in Appendix E.

5.1 Ablation Study

(1) Role of Modalities. We assess modality contribution through bimodal ablations, reported in

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 %.

Appendix B. Across all three datasets, full trimodal IAF remains the best configuration. The ablations are consistent with the empirical modality hierarchies identified by unimodal validation: removing text hurts most on MultiHuSE and UR-FUNNY, whereas on MUStARD the text-free A+V setting remains close to full trimodal performance, reflecting the stronger audio-visual signal.

(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 seen in Table 6, incorporating distilled representations 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 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 6: Distillation ablation: raw (Base) vs. distilled (Dist.) features within IAF. Δ = 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) shows that the learnable gate generally performs best, while the hierarchy gate remains competitive under strong modality imbalance. Appendix F further shows that mean gate weights remain distributed across modalities in all configurations, with no modality exceeding 0.53 on average.

6 Conclusion

We introduced Inverted Asymmetric Fusion (IAF), a two-stage multimodal framework for mitigating strong-modality collapse in discriminative fusion. By preserving the dominant modality through a PURE/ATTEND hierarchy and strengthening weaker modalities via knowledge distillation prior to fusion, IAF maintains the dominant pathway at its unimodal ceiling while improving fused performance in text-dominant and audio-visual-dominant settings. Across three benchmarks with different modality hierarchies, IAF consistently outperforms early, late, and symmetric fusion baselines, and surpasses prior published methods on UR-FUNNY and MUStARD. These findings suggest that protecting strong modalities from cross-modal interference is an important architectural principle for robust multimodal learning.

Limitations

This work has several limitations. First, IAF requires the dominant modality to be identified empirically through unimodal evaluation before fusion training. When modalities have similar predictive strength, the PURE/ATTEND assignment becomes less clear and the benefits of structural asymmetry may diminish. Moreover, the inferred hierarchy

depends on the encoder set: a stronger text encoder may appear dominant over a weaker video encoder even when the task itself relies more heavily on visual information. The hierarchy therefore reflects empirical encoder performance rather than intrinsic task structure. Although the dual-teacher design used for MUStARD partially addresses co-dominant settings, a more principled treatment remains open, for example through dynamic hierarchy assignment at the sample level.

Second, all experiments use frozen, pre-extracted features. This isolates the architectural contribution of IAF from encoder fine-tuning, but the framework has not been evaluated end to end, where joint optimisation could alter both the modality hierarchy and the effectiveness of distillation. In addition, our pathway analysis uses the original frozen unimodal classifier as a fixed probe, so it tests whether the dominant modality's decision boundary is preserved rather than whether it remains recoverable in principle. Future work could examine this more directly through probing or representational analyses such as Centred Kernel Alignment (CKA).

Finally, all datasets are English-only and confined to humor and sarcasm. It therefore remains unclear whether the modality hierarchies observed here, and the benefits of IAF, generalise to other languages or to tasks such as emotion recognition, visual question answering, or action recognition. Although the inclusion of E5-multilingual suggests potential cross-lingual applicability, this has not been evaluated, and cross-dataset generalisation remains 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.

Ronghao Lin and Haifeng Hu. 2024. Multi-Task Momentum Distillation for Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*, 15(2):549–565.

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 Jastrzę 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).

Yu Zhang, Bin Chen, Hongfei Ye, Zijian Gao, Tianjiao Wan, Long Lan, and Kele Xu. 2025. 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.

A Full Hyperparameter Configurations

Table 8 and Table 7 present the complete hyperparameter configurations for the distillation and fusion stages, across all datasets. 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	**Value**
KD loss weights (α, β, γ)	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	9×10^{-5}

Table 8: Distillation stage hyperparameters.

Hyperparameter	**MultiHuSE**	**UR-FUNNY**	**MUStARD**
Optimiser	AdamW	Adam	AdamW
Learning rate	6×10^{-5}	1×10^{-5}	1×10^{-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 7: Dataset-specific hyperparameters for the fusion

B Role of Modalities Ablation

We remove one modality at a time to quantify each modality's contribution to full trimodal performance. Table 9 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 $\approx$ 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.

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 9: Bimodal ablation of E5 and GloVe-based configuration for IAF

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 Table 9, confirming that the modality hierarchy findings are

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.

consistent across text encoders.

C Role of Regularisation Ablation

Table 11 reports the full regularisation ablation across all datasets and encoder configurations. BERT- and E5-based configurations are shown for MultiHuSE and MUStARD; results are directionally consistent across both encoders.

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 %.

D Role of Gate Mechanism Ablation

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 (≤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.

E Pathway Collapse: Visualisation

This section visualises pathway collapse in the distilled IAF and symmetric models (E5 + PE-Core, MultiHuSE) from two complementary perspectives: aggregate pathway accuracy across configurations and feature-space behaviour of the text pathway.

Figure 3 summarises dominant-modality pathway accuracy for all configurations in Table 5, comparing the unimodal ceiling, symmetric fusion, and IAF. Across all seven settings, symmetric fusion consistently reduces pathway accuracy, with a mean drop of 11.6 pp, whereas IAF matches the unimodal ceiling in every case. This pattern holds across datasets and encoder combinations, suggesting that strong-modality collapse is a structural property of symmetric cross-attention fusion rather than an artefact of any particular encoder.

Feature-space visualisation of collapse (t-SNE). Figure 4 shows a joint t-SNE embedding of the text-pathway features used at the final classification step, coloured by prediction correctness. Both

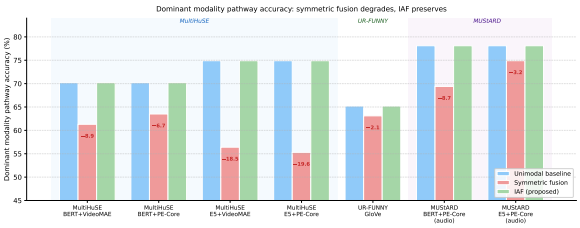

Figure 3: Dominant-modality pathway accuracy for all configurations (data from Table 5). Shaded regions group MultiHuSE (blue) and MUStARD (purple) configurations.

panels share a single embedding fitted on the combined feature matrix, so differences in error density reflect classifier behaviour rather than changes in the underlying feature geometry.

In the IAF panel (left), the frozen unimodal classifier ψ_T operates on the backbone features $\boldsymbol{h}^{(T)}$ on which it was trained; text-pathway accuracy on fold 4 is 76.7%. In the symmetric fusion panel (right), the shared classifier operates on post-cross-attention features $\tilde{\boldsymbol{h}}^{(T)}$; accuracy drops to 40.3%. The residual connection preserves a broadly similar manifold structure across both panels, yet the symmetric pathway produces substantially more errors in the same geometric space. This pattern is consistent with classifier mismatch: the shared head is not optimised for single-pathway inputs.

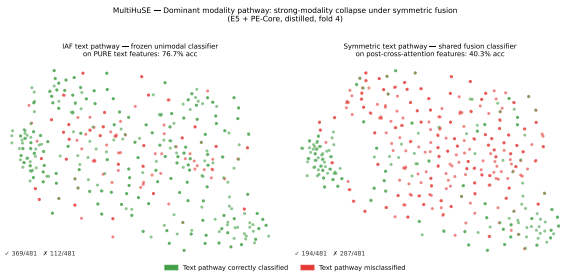

Figure 4: Joint t-SNE of features used at the final classification step for the text pathway under IAF (*left*) and symmetric fusion (*right*), fold 4 of MultiHuSE (E5 + PE-Core, distilled). Green = correctly classified by that pathway; red = misclassified.

F Gate Weight Analysis

The IAF gate network produces per-sample weights (w_T, w_A, w_V) that sum to one and determine each modality's contribution before classification. Table 13 reports mean gate weights averaged across cross-validation folds at the best checkpoint for all nine configurations. Two population-level patterns emerge. First, the modality assigned the PURE role consistently receives the highest mean weight: text leads on MultiHuSE ($\bar{w}_T \in [0.35, 0.47]$), whereas audio and video dominate on MUStARD ($\bar{w}_A + \bar{w}_V > 0.80$ in all configurations). On UR-FUNNY, weights are more evenly distributed ($\bar{w}_T = 0.372$, $\bar{w}_A = 0.310$, $\bar{w}_V = 0.318$), consistent with more balanced cross-modal cues in that corpus. Second, no modality exceeds a mean weight of 0.53 in any configuration, suggesting that IAF preserves multimodal integration rather than collapsing into a regularised unimodal model.

Text	**Feat.**	**Δ(%)**	$\bar{w}_T$	$\bar{w}_A$	$\bar{w}_V$
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.	+8.25	0.372	0.310	0.318
MUStARD *(Audio+Video* PURE | *Text* ATTEND*)*					
BERT+PE-Core	Base	$+3.47^{[B]}$	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

Table 13: Mean learned gate weights across CV folds at the best checkpoint. Δ = gain over dominant unimodal baseline (MultiHuSE: BERT = 68.26%, E5 = 74.86%; UR-FUNNY: GloVe = 62.47%; MUStARD: Dasheng = 78.12%). $^{[B]}/^{[E]}$: gap from BERT/E5 baseline. Base = raw pre-trained; Dist. = knowledge-distilled.

These averages, however, mask substantial within-dataset variation. Table 14 presents representative samples from the distilled IAF model (E5 + PE-Core) on MultiHuSE. Although the population average is $\bar{w}_T \approx 0.51$, individual samples range from near-exclusive text weighting ($w_T > 0.95$) to strong audio dominance ($w_A > 0.85$). Four cases illustrate this adaptive behaviour. In Category (i), unambiguous headline-style utterances labelled as neutral humour are classified primarily from text. In Category (ii), deadpan or context-dependent phrasing provides weak textual evidence, and the gate shifts weight toward audio and visual cues while preserving correct classification. In Category (iii), the model shows over-reliance on text: utterances such as *Age is of no importance unless you're a cheese* depend on delivery cues absent from the transcript, leading to misclassification. In Category (iv), audio becomes

dominant when the transcript is relatively uninformative; for example, *Tomorrow is 2-22-22. Happy Tuesday!* appears textually plain, but vocal warmth and intonation support affiliative classification, and the gate assigns about 90% of the weight to audio. Overall, the aggregate trends align with the PURE/ATTEND assignments, while the case studies show that the gate adapts to sample-level variation in modality reliability.

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 14: 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.

G Distillation Coefficient Sensitivity

Table 15 reports the effect of varying the three distillation loss coefficients (α: hard CE, β: soft KD, γ: feature alignment) on MultiHuSE. We evaluate three alternative configurations alongside the default setting: equal weights, a CE-heavy variant, and a soft-target-heavy variant. All distilled students are then used as inputs to IAF under the same fusion training procedure.

Fusion accuracy for IAF with E5, audio, and PE-Core varies by at most 1.20 percentage points across the four configurations. The two best-performing settings, the default and equal-weight variants, achieve the same top-line accuracy of 78.06%. CE-heavy weighting yields a small improvement in video distillation (+0.38 pp relative to the default) but reduces cross-modal soft transfer, leading to a 1.20 pp drop in fusion accuracy.

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

Table 15: Distillation coefficient sensitivity on MultiHuSE (5-fold CV, mean ± SD %). † = setting used in all main experiments. IAF columns use distilled students; Audio = Dasheng, Video = PE-Core throughout. "Range" reports max − min across the four settings.

The soft-target-heavy variant lowers video distillation performance (-0.83 pp relative to the default), suggesting that the weakest modality depends more on direct task supervision through CE than on soft targets from a teacher trained on a different modality. In this setting, increasing the weight on soft targets from the text teacher appears to overwhelm the video student's task-specific learning signal. Audio distillation remains stable across all settings, with a range of at most 0.33 pp, indicating that the Dasheng encoder adapts consistently regardless of coefficient choice. Overall, these results suggest that IAF is not highly sensitive to the precise coefficient values; the default setting was selected by manual tuning and then held fixed across all datasets.

H Knowledge Distillation Design Comparison

Table 16 situates MAKD relative to prior multimodal distillation approaches along five design axes: training stage, teacher source, objective, and optimisation procedure. The key distinction is that MAKD is a standalone pre-fusion stage that uses only the empirically dominant unimodal modality as teacher to strengthen weaker encoders before fusion, whereas prior multimodal KD methods typically operate within the fusion stage itself.

Method	Stage	Teacher	Purpose	Procedure
Cross-M KD (Albanie et al., 2018) (Thoker and 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 and Hu, 2024)	Within fusion	Unimodal + fused	Heterogeneity reduction	Concurrent
MAKD (Ours)	**Pre-fusion**	**Dom. unimodal**	**Encoder strengthening**	**Sequential**

Table 16: Design comparison of multimodal knowledge distillation methods. †Not designed in the context of a subsequent fusion stage.