Asymmetric Multimodal Learning with Residual Decomposition for Multimodal Intent Recognition

Yuxuan Zhai[1], Zongwei Li[1,*], Longfei Liu[2]
[1]Shanghai Institute of Technology
[2]State Key Laboratory for Novel Software Technology, Nanjing University
yux.zhai@gmail.com, lzw0118@sit.edu.cn, lf.liu@smail.nju.edu.cn

Abstract

Multimodal intent recognition predicts a speaker's intent from text, audio, and video, enabling conversational AI and human–computer interaction. However, existing methods share two limitations: (I) the three modalities are assigned equivalent roles even though text alone carries most of the predictive signal; (II) nonverbal cues that complement text are mixed with cues that merely repeat it. We propose Asymmetric Multimodal Learning with Residual Decomposition (AML), in which text produces the primary prediction and nonverbal signals enter only as a correction. The correction splits into Hierarchical Modality Calibration (HMC), which removes cues already present in text, and Asymmetric Multimodal Disentanglement (AMD), which carries cues that add new information. Extensive experiments on MIntRec2.0 and IEMOCAP-DA show that AML consistently surpasses prior state-of-the-art methods across all reported metrics, and additional analyses confirm the structural advantages of the asymmetric design.

1 Introduction

Multimodal intent recognition aims to utilise information from natural language together with nonverbal modalities such as speech audio and video so that machines can discern the communicative intent behind real-world utterances. It plays an increasingly important role in conversational AI, human–computer interaction (Xu, 2019), intelligent dialogue systems (Fan et al., 2022), and other human–robot interaction applications (Paul et al., 2022).

Prior works (Zhang et al., 2022, 2024) have pioneered this area by introducing large-scale multimodal intent datasets and adapting fusion strategies from multimodal sentiment analysis to construct initial baselines (Hazarika et al., 2020; Tsai et al., 2019; Rahman et al., 2020). Building on this foundation, one line of work strengthens the alignment between text and nonverbal streams through fine-grained interaction mechanisms, including token-level contrastive learning with modality-aware prompts (Zhou et al., 2024), shallow-to-deep interaction frameworks (Huang et al., 2024), and semantic relational reasoning over LLM-extracted fine-grained cues (Zhou et al., 2025). A second line couples fusion with an explicit out-of-distribution objective so that the model recognises unseen intents (Zhang et al., 2025). A third line draws on shared/private disentanglement (Bousmalis et al., 2016; Hazarika et al., 2020) and its recent counterfactual-intervention extension (Chen et al., 2024) to factor modality contributions before fusing them. Across all three lines, however, the resulting architectures share a common assumption (illustrated in Figure 1, left): text, audio, and video are peer contributors to a single fused representation, and the principal task of the model is to learn how to combine them in a balanced way.

Despite this progress, two critical issues remain unresolved across existing methods. First, since text alone already carries most of the predictive intent (Mullick et al., 2025), the additional engineering effort spent on the nonverbal modalities tends to add noise rather than additional information. Second, fusing all three modalities into one signal mixes complementary nonverbal cues with cues that echo text, leaving the two inseparable. Addressing both issues calls for a different architectural design rather than further engineering on existing mechanisms: one that (1) suppresses noise from the weaker modalities, and (2) disentangles complementary nonverbal evidence from cues that merely echo text.

These limitations expose a more fundamental gap in how the three modalities are integrated into the prediction. Even when explicit per-modality reliability weights are introduced, this only refines

*Zongwei Li is the corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 44630–44642
October 24-29, 2026 ©2026 Association for Computational Linguistics

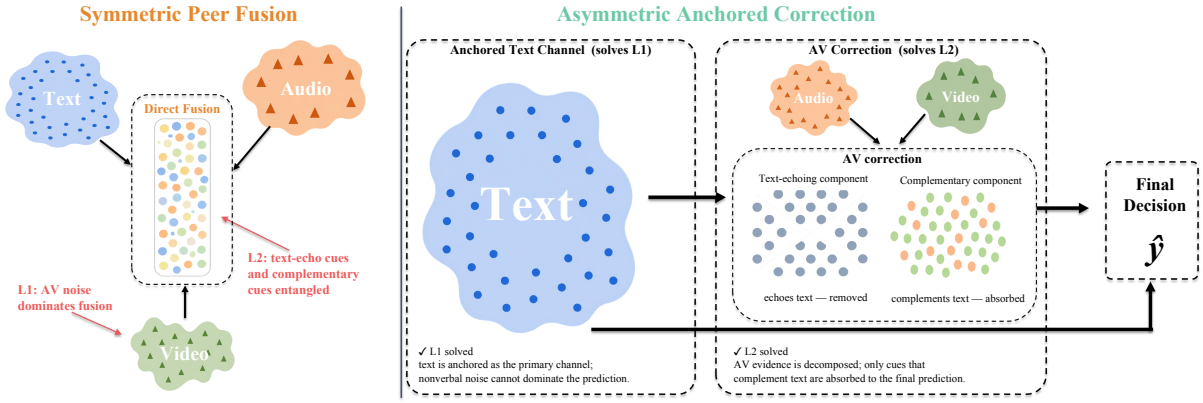


Figure 1: Conceptual comparison between vanilla multimodal fusion and AML.

how loudly each modality is allowed to speak, leaving open the more basic question of what role each modality plays in shaping the prediction. In our approach, text serves as the primary source of the prediction and the nonverbal modalities only refine it, so that each modality contributes in proportion to the information it carries. In addition to this, only nonverbal cues that genuinely supplement the text are allowed to influence the prediction; cues that merely echo text are removed.

To address the first of these issues, we introduce AML (Asymmetric Multimodal Learning). Given that text already carries the bulk of the predictive intent, AML adopts an asymmetric architecture. Text is established as the primary predictive channel, while the nonverbal modalities are restricted to a complementary refining role. The text representation is initially trained in isolation to consolidate its role as the primary predictor before any nonverbal information enters the architecture. Subsequently, with the text channel frozen, a separate branch is trained to combine it with the nonverbal modalities and yield a structurally separate correction that resides outside the text prediction pathway. This structural separation constrains the nonverbal contribution to a supplementary role, preserving the original text prediction as the baseline. By construction, disabling the correction exactly recovers the text-only prediction. The correction can nevertheless hurt an individual prediction when its direction is wrong; improvements over the text-only channel are therefore empirical aggregate results, not per-sample guarantees.

To address the second of these issues, we structurally separate the multimodal correction into two parts that handle echo cues and complementary contributions differently. We introduce Hierarchical Modality Calibration (HMC), a per-sample scaling of the text prediction that removes echo cues from the class decision. Complementary contributions, by contrast, flow through Asymmetric Multimodal Disentanglement (AMD), the only channel that absorbs nonverbal evidence into the final prediction. Consequently, only complementary nonverbal cues can alter the final prediction, while cues that merely echo text are removed before they reach it. Together, HMC and AMD achieve a structural disentanglement that makes the architecture robust against misleading or redundant nonverbal signals.

Beyond raw accuracy, controlled axis-specific interventions on the nonverbal correction localise the gain to its *direction* — which alternative class the nonverbal cues propose — while its *magnitude* — how strongly the cues vote against text — is largely fungible above a low threshold. The corresponding intervention on the text-aligned scalar is statistically inert, the predicted outcome of the architectural identity.

Our contributions are as follows.

- We introduce a new training framework for multimodal intent recognition, where text serves as the primary prediction channel and the nonverbal modalities provide auxiliary refinement, so each modality contributes through its proper role rather than as equals.
- We propose AML (Asymmetric Multimodal Learning with Residual Decomposition), an architecture that achieves structural robustness against misleading or redundant nonverbal cues, delivering reliable intent recognition across diverse multimodal input qualities.
- AML achieves state-of-the-art results on two multimodal intent recognition benchmarks. Ad-

ditional analyses verify its structural advantages.

2 Related Work

2.1 Multimodal Intent Recognition

Multimodal intent recognition predicts a speaker's communicative intent from paired transcript, audio, and video. The MIntRec (Zhang et al., 2022) and MIntRec2.0 (Zhang et al., 2024) benchmarks, together with dialogue-act corpora built on top of IEMOCAP (Busso et al., 2008), anchor evaluation in this trimodal setting. Early baselines port architectures from multimodal sentiment analysis. MAG-BERT (Rahman et al., 2020) inserts a gating module that shifts BERT hidden states with acoustic and visual offsets, while MulT (Tsai et al., 2019) couples unaligned modality streams via directional pairwise cross-attention. Building on these, recent work refines text–nonverbal interaction at the architectural level, where TCL-MAP (Zhou et al., 2024) combines modality-aware prompting with token-level contrastive learning on label tokens and SDIF-DA (Huang et al., 2024) stacks shallow-to-deep interaction blocks with data augmentation. MIntOOD (Zhang et al., 2025) extends intent classification to out-of-distribution detection. Beyond in-network designs, LGSRR (Zhou et al., 2025) delegates semantic extraction to a large language model and reasons relationally over the resulting cues. Across these designs, Mullick et al. (2025) nonetheless find that a text-only large language model (Mistral-7B) outperforms competitive multimodal fusion models on these benchmarks, attributing the gap to strong textual bias in the data.

2.2 Modality Imbalance

Jointly trained multimodal classifiers often underuse their weaker modalities. Gradient Blending (Wang et al., 2020) observes that a multimodal network can generalise worse than the best of its unimodal sub-networks, and blends gradients from joint and unimodal losses to balance the optimisation. OGM (Peng et al., 2022) attributes the gap to gradient suppression by the dominant modality and applies on-the-fly gradient modulation that rebalances per-modality update magnitudes during training. Du et al. (2023) show that joint multimodal training degrades the unimodal features each branch learns, and propose to decouple unimodal feature learning from the multimodal interaction stage.

2.3 Residual Learning

Residual learning decomposes a model's output into an anchored baseline and a learnable correction, so that the trainable components only model the deviation from the baseline rather than re-derive the full signal. He et al. (2016) introduce this pattern in deep residual networks through identity shortcuts added to the output of each learnable block. In transfer learning, residual adapters apply the same contract at the parameter level. Rebuffi et al. (2017) attach small additive modules to a frozen vision backbone, and Houlsby et al. (2019) extend the design to language models through the adapter framework, in which each adapter is initialised close to identity so that the frozen baseline is preserved at the start of training.

3 Method

AML comprises two training stages (Figure 2). The first stage trains a text-anchored primary classification channel; the second stage freezes the text channel and learns a multimodal correction decomposed orthogonally against the text-logit direction.

3.1 Problem Formulation

The task of multimodal intent recognition is to predict the communicative intent of a speaker from paired text, audio, and video. A multimodal intent recognition dataset $\mathcal{D} = \{(x_i^{\mathrm{T}}, x_i^{\mathrm{A}}, x_i^{\mathrm{V}}, y_i)\}_{i=1}^{N}$ provides, for each utterance i, a transcript x^{T}, an audio waveform x^{A}, a video clip x^{V}, and an intent label $y \in \{1, \ldots, K\}$, where K is the number of intent classes defined by the benchmark. The model receives the three modalities jointly and must produce a single intent prediction $\hat{y} \in \{1, \ldots, K\}$. Each input modality is encoded into a fixed-dimensional feature vector through a modality-specific pretrained encoder,

$$r^{\mathrm{T}} = f_{\mathrm{T}}(x^{\mathrm{T}}), \quad r^{\mathrm{A}} = f_{\mathrm{A}}(x^{\mathrm{A}}), \quad r^{\mathrm{V}} = f_{\mathrm{V}}(x^{\mathrm{V}}), \tag{1}$$

with each encoder followed by mean pooling along the time dimension; we denote the resulting text, audio, and video features as $r^{\mathrm{T}}, r^{\mathrm{A}}, r^{\mathrm{V}} \in \mathbb{R}^d$ with a common feature dimension d.

3.2 Residual Geometry Structuring (RGS)

The first stage trains the text head in isolation so that it provides a stable predictive direction for the orthogonal decomposition introduced in Section 3.3. We attach two heads to the text representation r^{T}, a classifier head and a non-negative

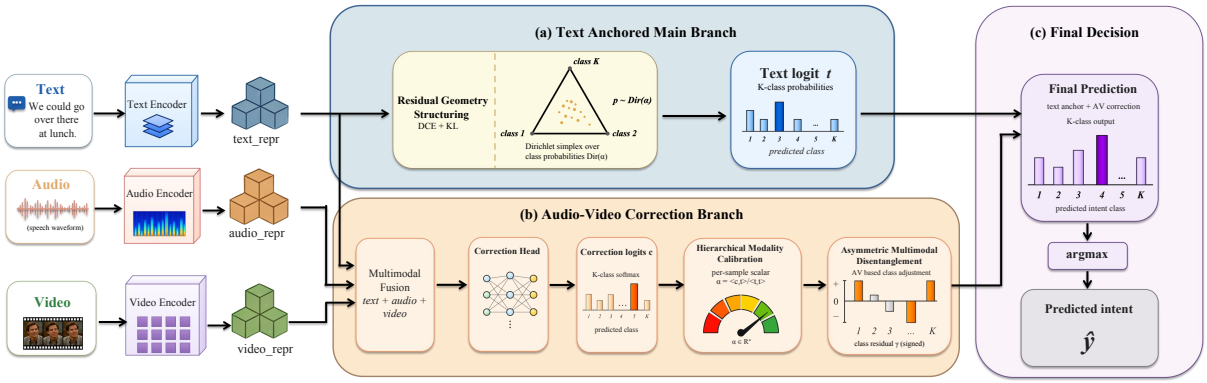


Figure 2: Overview of the proposed AML framework.

evidence head,

$$t = f_{\text{cls}}(r^{\text{T}}) \in \mathbb{R}^K, \qquad e = f_{\text{evi}}(r^{\text{T}}) \in \mathbb{R}^K_{\geq 0}, \tag{2}$$

where t is used downstream as the text prediction direction and e supplies evidence for an associated Dirichlet posterior. The evidence vector parameterises a Dirichlet posterior over the class probability simplex,

$$\text{Dir}(\boldsymbol{\alpha}_{\text{dir}}), \qquad \boldsymbol{\alpha}_{\text{dir}} = e + \mathbf{1}, \tag{3}$$

following the evidential deep learning framework of Şensoy et al. (2018). Treating the prediction as a distribution over the simplex rather than as a point estimate encourages the text head to learn a calibrated confidence in addition to a correct $\arg\max$, which is what later allows the multimodal correction to be decomposed cleanly against the text direction.

We train both heads jointly with a cross-entropy loss on the text logits and an evidential objective. The latter combines a Dirichlet cross-entropy (DCE) on the expected log-likelihood under $\text{Dir}(\boldsymbol{\alpha}_{\text{dir}})$ (Şensoy et al., 2018) with a Kullback–Leibler regulariser on off-target evidence (Han et al., 2021b):

$$\begin{aligned}\mathcal{L}_{\text{P1}} &= \mathcal{L}_{\text{CE}}(t, y) \\ &\quad + \lambda_{\text{evi}}[\mathcal{L}_{\text{DCE}}(\boldsymbol{\alpha}_{\text{dir}}, y) + \lambda_{\text{KL}}\, \mathcal{L}_{\text{KL}}]\,.\end{aligned} \tag{4}$$

Here $\mathcal{L}_{\text{KL}} = \text{KL}\big(\text{Dir}(\tilde{\boldsymbol{\alpha}}_{\text{dir}}) \,\|\, \text{Dir}(\mathbf{1})\big)$. The Dirichlet cross-entropy term takes the explicit form

$$\mathcal{L}_{\text{DCE}}(\boldsymbol{\alpha}_{\text{dir}}, y) = \sum_{k=1}^{K} y_k \left(\psi(S) - \psi(\alpha_{\text{dir},k})\right), \tag{5}$$

where $S = \sum_k \alpha_{\text{dir},k}$ is the Dirichlet strength and $\psi(\cdot)$ is the digamma function, and the KL regulariser uses the flattened parameters $\tilde{\boldsymbol{\alpha}}_{\text{dir}} = y + (1 - y) \odot \boldsymbol{\alpha}_{\text{dir}}$, which set the target-class concentration to one and retain the off-target concentrations for KL penalisation.

3.3 Orthogonal Residual Decomposition

The text head is frozen after Phase 1 and supplies a stable predictive direction in the logit space. To incorporate audio and video without disturbing this direction, we first fuse the three modality representations through a multimodal fusion MLP,

$$r^{\text{F}} = \text{MLP}_{\text{fuse}}\big([\, r^{\text{T}}; r^{\text{A}}; r^{\text{V}}\,]\big) \in \mathbb{R}^{512}, \tag{6}$$

where $[\cdot;\cdot;\cdot]$ denotes channel-wise concatenation, and then map the fused feature through a correction head f_{corr} to a correction logit vector,

$$c = f_{\text{corr}}(r^{\text{F}}) \in \mathbb{R}^K, \tag{7}$$

intended to adjust the text prediction with cross-modal evidence. An unconstrained correction would in general rewrite the text head's prediction; we instead decompose c explicitly so that the text-aligned and the text-orthogonal contributions of the correction can be handled separately.

Because c contains both information that overlaps with the text prediction direction and information that is orthogonal to it, we decompose c explicitly against the frozen text logits t:

$$c = \alpha\, t + \gamma, \qquad \langle \gamma, t \rangle = 0. \tag{8}$$

The decomposition is unique once t is fixed (Appendix B provides the full derivation). The two components, named below, capture functionally distinct aspects of the correction. Our use of *echo* and *complementary* is functional rather than semantic: the text-aligned component is called echo because, under the empirically non-negative scaling reported in Section 4.5, it preserves the text prediction's class ranking. The orthogonal residual is called complementary because it is the only component with the freedom to change the prediction

relative to the text head. This functional guarantee does not establish semantic disentanglement of linguistically redundant and complementary cues.

Hierarchical Modality Calibration. The scalar α quantifies how much of the multimodal correction lies along the text head's prediction direction. It is defined, for non-zero t, as the projection of c onto t,

$$\alpha = \frac{\langle c, t \rangle}{\langle t, t \rangle}, \tag{9}$$

so that $\alpha\, t$ is the text-aligned component of c. Because α is determined entirely by inner products of c and t at inference, it introduces no additional learnable parameters into the architecture: the calibration is read directly off the correction head's output rather than learned by a separate scoring module. Operationally, α calibrates the modality hierarchy of the architecture: when $\alpha > 0$, the text-aligned component is a positive rescaling of t that preserves the text head's $\arg\max$, and any predictive shift the correction induces must be carried by the orthogonal residual γ defined next.

The frozen text channel and the multimodal correction are combined through a learned mixing scalar $\beta(x) \in (0, 1)$,

$$\text{final}(x) \;=\; \beta(x)\, t \;+\; \big(1 - \beta(x)\big)\, c, \tag{10}$$

and the predicted class is the index of the largest logit,

$$\hat{y} \;=\; \arg \max_{k \in \{1,\ldots,K\}} \text{final}(x)_k. \tag{11}$$

Asymmetric Multimodal Disentanglement. The residual γ carries the portion of the multimodal correction that is uncorrelated with the text head's prediction direction. It is obtained by subtracting the text-aligned component from c,

$$\gamma \;=\; c - \alpha\, t, \tag{12}$$

and is orthogonal to t by construction. Substituting Equation 12 into the final prediction (Equation 10) and collecting the text-aligned terms yields

$$\text{final}(x) \;=\; \big[\beta + (1-\beta)\alpha\big]\, t \;+\; (1-\beta)\, \gamma, \tag{13}$$

in which the bracketed coefficient is positive whenever $\alpha > 0$; the text-aligned term is then argmax-invariant under positive rescaling, and γ is the sole component of the architecture capable of altering the predicted class. Because γ depends on the full correction c rather than on α alone, different AV inputs paired with the same text produce different residuals, so the prediction-shifting signal varies on a per-sample basis even when the text head's output is held fixed. At inference the model performs a single forward pass through the text channel, the audio and video encoders, the fusion MLP, and the correction head, and emits $\hat{y}$ directly from Equation 11.

3.4 Training Objective

For a mini-batch $\mathcal{B}$, the orthogonality regulariser uses the detached batch-level residual

$$\begin{aligned} \widehat{\alpha} &= \text{stopgrad}\left(\frac{\sum_{i \in \mathcal{B}} \langle c_i, t_i \rangle}{\sum_{i \in \mathcal{B}} \langle t_i, t_i \rangle} \right), \\ \widehat{\gamma}_i &= c_i - \widehat{\alpha} t_i. \end{aligned} \tag{14}$$

Although its batch-summed projection is zero, individual $\langle \widehat{\gamma}_i, t_i \rangle$ need not vanish, so the per-sample squared cosines are non-degenerate. The exact α_i and γ_i in Equations 9–12 remain the inference-time decomposition; only the training regulariser uses $\widehat{\alpha}$ and $\widehat{\gamma}_i$.

Phase 2 is trained with the mini-batch composite loss

$$\begin{aligned} \mathcal{L}_{\text{P2}} = \mathcal{L}_{\text{CE}}(\text{final}, y) \\ + \frac{\lambda_{\text{orth}}}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \, \cos^2(\widehat{\gamma}_i, t_i) \\ + \frac{\lambda_{\text{para}}}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \|\widehat{\gamma}_{\parallel, i}\|^2 \\ + \lambda_{\text{aux}} \big(\mathcal{L}^{\text{A}}_{\text{CE}} + \mathcal{L}^{\text{V}}_{\text{CE}} \big), \end{aligned} \tag{15}$$

where the regulariser quantities are

$$\begin{aligned} w_i &= 1 - \frac{H(\text{softmax}(t_i))}{\log K}, \\ \cos^2(\widehat{\gamma}_i, t_i) &= \frac{\langle \widehat{\gamma}_i, t_i \rangle^2}{\max(\|\widehat{\gamma}_i\|, \varepsilon)^2 \, \max(\|t_i\|, \varepsilon)^2}, \\ \|\widehat{\gamma}_{\parallel, i}\|^2 &= \frac{\langle \widehat{\gamma}_i, t_i \rangle^2}{\max(\langle t_i, t_i \rangle, \varepsilon)}. \end{aligned} \tag{16}$$

Here, $\varepsilon = 10^{-8}$ safeguards zero norms. The parallel-magnitude penalty discourages text-aligned rescaling, which is argmax-invariant under the empirically non-negative alignment coefficients; all regulariser weights are selected on the development set. The joint cross-entropy is

$$\mathcal{L}_{\text{CE}}(\text{final}, y) = -\sum_{k=1}^{K} y_k \, \log \, \text{softmax}(\text{final})_k, \tag{17}$$

Methods	MIntRec2.0						IEMOCAP-DA					
	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)	ACC (↑)	F1 (↑)	P (↑)	R (↑)	WF1 (↑)	WP (↑)
MISA	55.16	49.51	51.80	49.92	55.05	57.06	73.76	72.26	73.03	72.51	73.59	73.87
MAG-BERT	60.38	54.74	57.51	54.54	59.61	60.00	74.25	72.07	73.18	72.33	74.03	74.28
MulT	60.66	54.12	58.02	53.77	59.55	60.12	73.74	72.28	73.40	72.21	73.66	74.15
TCL-MAP	58.24	52.25	54.28	52.41	57.24	57.55	74.37	72.63	74.02	72.39	74.21	74.76
SDIF-DA	58.06	51.95	53.17	52.16	57.47	57.85	74.19	72.34	73.76	72.39	74.04	74.77
MIntOOD	58.25	51.73	56.79	50.99	57.11	58.65	74.56	71.31	72.70	70.89	74.40	74.65
LGSRR	60.46	55.35	59.33	55.09	59.72	60.85	74.95	72.99	74.27	72.74	74.88	75.47
AML	**62.74**	**58.26**	**61.87**	**57.35**	**62.06**	**62.60**	**75.32**	**73.41**	**74.69**	**77.14**	**75.26**	**75.66**

Table 1: Main results on MIntRec2.0 and IEMOCAP-DA.

Methods	In-scope Classification				In-scope + Out-of-scope Classification			
	ACC(%)	WF1(%)	WP(%)	R(%)	F1-IS(%)	ACC(%)	F1-OOS(%)	F1(%)
w/o Out-of-scope samples								
TEXT	59.30	58.01	58.85	51.31	43.37	43.24	30.40	42.96
MAG-BERT	60.58	59.68	59.98	55.10	46.48	44.80	34.03	46.08
MulT	60.66	59.55	60.12	53.77	45.65	46.14	38.57	45.42
AML	**62.74**	**62.06**	**62.60**	**57.35**	**48.74**	**49.96**	**45.04**	**48.62**
w/ Out-of-scope samples								
TEXT	59.99	58.62	58.65	52.11	45.83	55.61	61.54	46.34
MAG-BERT	60.12	59.11	58.83	53.79	47.52	56.20	62.47	48.00
MulT	60.18	58.82	59.38	52.56	46.88	56.00	61.66	47.35
AML	**61.08**	**60.06**	**60.83**	**54.76**	**48.39**	**57.12**	**63.56**	**48.88**

Table 2: Main results on MIntRec2.0-OOD.

and the auxiliary losses on the audio and video encoders take the same cross-entropy form on their respective heads,

$$\mathcal{L}_{\mathrm{CE}}^{\mathrm{A}} = \mathcal{L}_{\mathrm{CE}}\big(f_{\mathrm{A,head}}(r^{\mathrm{A}}), y\big), \\ \mathcal{L}_{\mathrm{CE}}^{\mathrm{V}} = \mathcal{L}_{\mathrm{CE}}\big(f_{\mathrm{V,head}}(r^{\mathrm{V}}), y\big). \tag{18}$$

4 Experiments

4.1 Experimental Setting

Datasets and Evaluation Metrics. We conduct experiments on two multimodal intent recognition benchmarks, MIntRec2.0 (Zhang et al., 2024) and IEMOCAP-DA (Busso et al., 2008; Zhou et al., 2025), both of which target multimodal intent recognition; MIntRec2.0 additionally provides an out-of-scope test partition for joint in-scope classification and OOD detection. We adopt six metrics to evaluate in-scope classification: accuracy (ACC), macro-F1 (F1), macro-precision (P), macro-recall (R), weighted-F1 (WF1), and weighted-precision (WP). To evaluate out-of-scope classification, we utilize accuracy (ACC), macro F1-score over all classes (F1), in-scope F1 (F1-IS), and out-of-scope F1 (F1-OOS).

Implementation Details. For feature extraction, we adopt BERT-large-uncased (Devlin et al., 2019) for text, WavLM-base (Chen et al., 2022) for audio, and a Swin-Transformer ROI extractor (Liu et al., 2021) for video. The composite loss is optimised with AdamW on a single NVIDIA RTX 5090 GPU, and reported numbers are averaged across five training runs. Further implementation and dataset details are provided in Appendix A.

4.2 Comparative Baselines

For comparison, we consider the following multimodal intent recognition methods as baselines: (1) MAG-BERT (Rahman et al., 2020); (2) MulT (Tsai et al., 2019); (3) MISA (Hazarika et al., 2020); (4) TCL-MAP (Zhou et al., 2024); (5) SDIF-DA (Huang et al., 2024); (6) MIntOOD (Zhang et al., 2025); (7) LGSRR (Zhou et al., 2025).

4.3 Main Results

Results on MIntRec2.0 and IEMOCAP-DA. Table 1 presents the comparison of AML against all baselines. Our method achieves the best results across all six evaluation metrics on both benchmarks. On MIntRec2.0, AML outperforms the

Ablation	MIntRec2.0						IEMOCAP-DA					
	ACC(↑)	F1(↑)	P(↑)	R(↑)	WF1(↑)	WP(↑)	ACC(↑)	F1(↑)	P(↑)	R(↑)	WF1(↑)	WP(↑)
w/o RGS	61.65	55.87	59.24	55.05	60.66	60.93	74.15	71.07	69.44	76.20	74.03	74.55
w/o HMC	61.78	56.81	60.94	55.98	60.91	61.66	74.01	71.44	70.27	76.97	73.93	74.37
w/o AMD mag	61.19	55.72	60.27	54.92	60.22	61.09	73.90	70.93	70.67	74.78	73.78	74.12
w/o AMD dir	58.26	52.55	55.25	51.96	57.23	57.65	73.27	70.32	70.47	74.91	73.10	73.44
w/o correction	59.32	53.77	57.76	53.11	58.33	59.36	73.13	70.30	71.65	74.38	72.97	73.78
w/o text	60.30	54.41	58.53	53.55	59.54	60.69	74.12	69.70	69.14	74.18	74.04	75.35
AML	**62.74**	**58.26**	**61.87**	**57.35**	**62.06**	**62.60**	**75.32**	**73.41**	**74.69**	**77.14**	**75.26**	**75.66**

Table 3: Ablation studies on the MIntRec2.0 and IEMOCAP-DA datasets.

previous state-of-the-art LGSRR by 2.28 percentage points on ACC, 2.91 points on F1, and 2.34 points on WF1, confirming that the text-anchored asymmetric architecture effectively leverages nonverbal evidence for fine-grained intent recognition. AML also achieves consistent improvements on IEMOCAP-DA, demonstrating that the asymmetric design generalises across tasks of different granularity.

Results on MIntRec2.0-OOD. Table 2 presents the out-of-scope detection results on MIntRec2.0 under two settings: training without out-of-scope samples and training with outlier exposure (Hendrycks et al., 2019). AML achieves the best performance across all evaluation columns in both settings, confirming that the asymmetric architecture extends naturally from in-scope classification to joint out-of-scope detection.

4.4 Ablation Study

To better understand how each component of AML contributes to its overall performance, we conduct ablation experiments and report the results in Table 3.

Effect of Residual Geometry Structuring. We replace the Dirichlet evidential supervision with a standard cross-entropy objective, which decreases F1 by 2.39 points. This change is attributed to the geometric role of evidential training: it structures the text logits in a form that supports downstream correction, whereas vanilla cross-entropy fails to induce the same orthogonal geometry. When the evidential supervision is active, the correction channel cleanly separates complementary nonverbal evidence from redundant content that the text channel already captures.

Effect of Hierarchical Modality Calibration. We freeze the Hierarchical Modality Calibration value at its empirical mean throughout training, which leads to a 0.96-point decrease in ACC. The decrease proves that the calibration provides essential gradient guidance during training: it aligns the multimodal correction with the text channel's prediction direction for each utterance, which a fixed value cannot reproduce.

Effect of Asymmetric Multimodal Disentanglement. Asymmetric Multimodal Disentanglement is the channel through which nonverbal evidence enters the final prediction. As shown in Table 3, restricting its magnitude yields a moderate drop, whereas permuting its direction yields a substantially larger drop on both datasets. Without a reliable direction, the multimodal correction can no longer point the prediction to the correct alternative class, indicating that direction is what carries the nonverbal class signal. The multimodal correction therefore identifies which alternative class to propose when the text channel is uncertain, rather than reinforcing the existing text prediction.

Effect of Removing the Text Channel. In the full architecture, the text prediction provides the shared reference frame that enables RGS, HMC, and AMD to function as a coherent decomposition system. Without this reference frame, the nonverbal variant still reaches 60.30 ACC on MIntRec2.0 and 74.12 on IEMOCAP-DA, confirming that the nonverbal features carry genuine discriminative capacity. The full AML substantially surpasses both this variant and the text-only variant on all metrics, confirming that the asymmetric decomposition generates value beyond what either modality provides on its own.

4.5 Analysis of the Per-Sample Signal

When a prediction shifts away from what text alone would predict, which component of the architecture is actually responsible? Three components are candidates: the text channel itself, the text-aligned component inside the nonverbal correction, and the

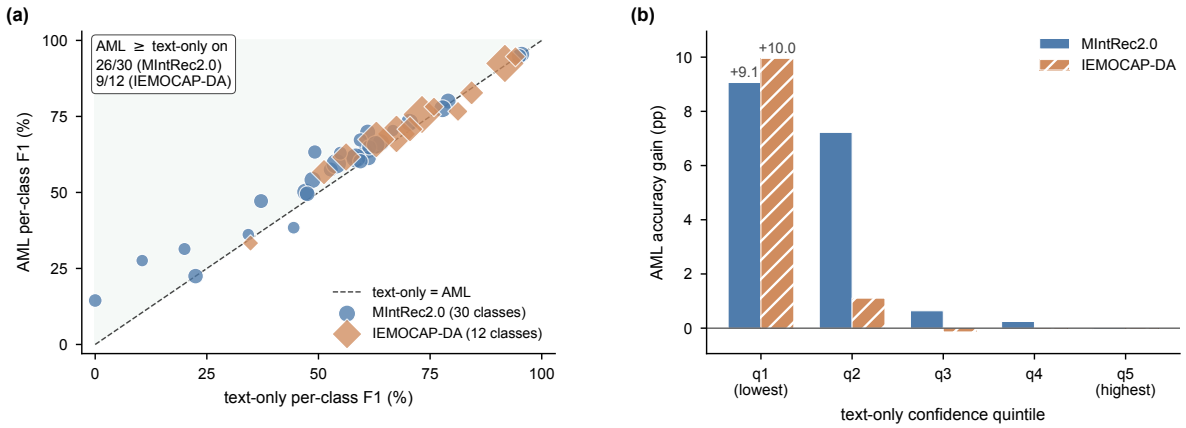


Figure 3: Class-level and sample-level effect of the nonverbal correction.

orthogonal residual that lies outside the text channel's span. An algebraic answer is available before any data is collected: the text-alignment coefficient is empirically non-negative across every sample in both datasets, so the multiplier in front of the text channel is a strictly positive scalar that cannot alter the predicted class. Only the orthogonal residual has the algebraic freedom to change it.

Two matched tests confirm this prediction. When the mixture weight is set so that the correction is disabled, the resulting prediction matches the text-only floor utterance-by-utterance across every seed on both datasets, not merely on average. A per-sample shuffle of the orthogonal residual at evaluation then drops accuracy substantially on both benchmarks, while the identical permutation applied to the text-alignment coefficient is statistically inert. The three views converge on the same architectural site: the orthogonal residual is what moves predictions; the text channel is anchored and its scaling coefficient is structurally inert at inference.

4.6 Geometric Decomposition of the Correction

With the per-sample signal localised to the orthogonal residual, we ask what the residual itself contains. The residual admits a further decomposition along two geometric axes: a magnitude axis that encodes how strongly the nonverbal features disagree with text, and a direction axis that encodes which alternative class they propose. We ablate the two axes independently in Table 3. Restricting the magnitude axis at training time causes a moderate accuracy drop on both benchmarks. Destroying the direction axis at evaluation by a per-sample shuffle drops accuracy substantially more: roughly three times the magnitude effect on MIntRec2.0 and clearly larger on IEMOCAP-DA. The nonverbal correction therefore behaves as a categorical disambiguator — deciding which alternative class to propose when text is uncertain — rather than as a confidence amplifier of an already correct text prediction. The latter role is precluded by construction: the text-aligned component is argmax-invariant.

4.7 Fine-grained Analysis of the Correction's Effect

The asymmetric design predicts that the nonverbal correction should make a measurable difference only where the text channel alone is already weak. We test this prediction along two complementary slices in Figure 3. At the class level, we plot per-class F1 of AML against per-class F1 of the text-only fall-back; AML matches or improves text-only on 26 of 30 intent classes on MIntRec2.0 and on 9 of 12 dialogue acts on IEMOCAP-DA, with the largest improvements concentrated on the classes that are hardest for text alone. At the utterance level, we stratify the test set by text-only top-1 confidence and report the AML accuracy gain within each quintile. The gain is heavily concentrated in the lowest-confidence quintile on both benchmarks and generally decreases to near zero on the highest-confidence utterances. The correction's aggregate gain is therefore selective: it is largest where text is uncertain and has no net accuracy gain in the highest-confidence quintile. This is the behaviour the asymmetric architecture targets.

5 Conclusion

In this paper, we propose AML (Asymmetric Multimodal Learning) for multimodal intent recognition. AML positions text as the primary prediction channel and nonverbal modalities as auxiliary correction, providing a principled resolution to the tension between exploiting multimodal evidence and preserving the empirical dominance of text. This asymmetric design restricts the nonverbal modality to contributing complementary class information rather than reinforcing the existing text prediction, offering a sharper characterisation of the nonverbal role than symmetric fusion permits. Our work not only demonstrates consistent state-of-the-art performance on two multimodal intent recognition benchmarks, but also carries significant implications for asymmetric architecture design wherever one modality structurally dominates. We further validate the generality of the asymmetric-fusion design on multimodal sentiment analysis benchmarks in Appendix C.

Limitations

Despite the consistent improvements reported in this work, several open questions remain for future investigation. First, the benchmarks used in our evaluation are all text-dominant: text is empirically the strongest modality for predicting the label, and our asymmetric framing is instantiated around that structure. How an analogous asymmetric architecture should be re-instantiated for tasks with a different dominance pattern remains an open question that we leave to future research. Second, the proposed asymmetric framework formalises the interaction between text and the nonverbal correction in a controlled geometric manner. However, it does not fully account for the subtle and context-dependent pragmatic cues that shape intent in real-world conversations, where nonverbal evidence carries substantially more noise than in curated benchmarks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 72474135).

References

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Proc. NeurIPS*. https://proceedings.neurips.cc/paper/2016/hash/45fbc6d3e05ebd93369ce542e8f2322d-Abstract.html.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359. https://link.springer.com/article/10.1007/s10579-008-9076-6.

Sanyuan Chen, Chengyi Wang, Zhuo Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113.

Zhanpeng Chen, Zhihong Zhu, Xianwei Zhuang, Zhiqi Huang, and Yuexian Zou. 2024. Dual-oriented disentangled network with counterfactual intervention for multimodal intent detection. In *Proc. EMNLP*. https://doi.org/10.18653/v1/2024.emnlp-main.972.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*. https://doi.org/10.18653/v1/N19-1423.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. 2023. On uni-modal feature learning in supervised multi-modal learning. In *Proc. ICML*. https://proceedings.mlr.press/v202/du23e.html.

Yan Fan, Chengyu Wang, Peng He, and Yunhua Hu. 2022. Building multi-turn query interpreters for E-Commercial chatbots with sparse-to-dense attentive modeling. In *Proc. WSDM*, pages 1577–1580. https://doi.org/10.1145/3488560.3502189.

Wei Han, Hui Chen, and Soujanya Poria. 2021a. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proc. EMNLP*. https://doi.org/10.18653/v1/2021.emnlp-main.723.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2021b. Trusted multi-view classification. In *Proc. ICLR*. https://openreview.net/forum?id=OOsR8BzCnl5.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proc. ACM MM*. https://doi.org/10.1145/3394171.3413678.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. CVPR*. https://doi.org/10.1109/CVPR.2016.90.

Xilin He, Haijian Liang, Boyi Peng, Weicheng Xie, Muhammad Haris Khan, Siyang Song, and Zitong Yu. 2025. MSAmba: Exploring multimodal sentiment analysis with state space models. In *Proc. AAAI*. https://ojs.aaai.org/index.php/AAAI/article/view/32120.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. 2019. Deep anomaly detection with outlier exposure. In *Proc. ICLR*. https://openreview.net/forum?id=HyxCxhRcY7.

Neil Houlsby, Andrei Giurgiu, Stanisław Jastrzębski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proc. ICML*. https://proceedings.mlr.press/v97/houlsby19a.html.

Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. 2024. SDIF-DA: A shallow-to-deep interaction framework with data augmentation for multi-modal intent detection. In *Proc. ICASSP*. https://doi.org/10.1109/ICASSP48485.2024.10446922.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proc. ICCV*. https://doi.org/10.1109/ICCV48922.2021.00986.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-SENA: An integrated platform for multimodal sentiment analysis. In *Proc. ACL Demo*. https://doi.org/10.18653/v1/2022.acl-demo.12.

Ankan Mullick, Saransh Sharma, Abhik Jana, and Pawan Goyal. 2025. Text takes over: A study of modality bias in multimodal intent detection. In *Proc. EMNLP (Main)*. https://doi.org/10.18653/v1/2025.emnlp-main.1226.

Sheuli Paul, Michael Sintek, Veton Këpuska, Marius Silaghi, and Liam Robertson. 2022. Intent based multimodal speech and gesture fusion for human–robot communication in assembly situation. In *Proc. ICMLA*, pages 760–763. https://doi.org/10.1109/ICMLA55696.2022.00127.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proc. CVPR*. https://doi.org/10.1109/CVPR52688.2022.00806.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proc. ACL*. https://doi.org/10.18653/v1/2020.acl-main.214.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Proc. NeurIPS*. https://proceedings.neurips.cc/paper/2017/hash/e7b24b112a44fdd9ee93bdf998c6ca0e-Abstract.html.

Murat Şensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Proc. NeurIPS*. https://proceedings.neurips.cc/paper_files/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proc. ACL*. https://doi.org/10.18653/v1/P19-1656.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proc. CVPR*. https://doi.org/10.1109/CVPR42600.2020.01271.

Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. 2025. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proc. AAAI*. https://doi.org/10.1609/aaai.v39i2.32152.

Zhuojia Wu, Qi Zhang, Duoqian Miao, Kun Yi, Wei Fan, and Liang Hu. 2024. HyDiscGAN: A hybrid distributed cGAN for audio-visual privacy preservation in multimodal sentiment analysis. In *Proc. IJCAI*, pages 6550–6558. https://doi.org/10.24963/ijcai.2024/724.

Wei Xu. 2019. Toward human-centered AI: A perspective from human–computer interaction. *Interactions*, 26(4):42–46. https://doi.org/10.1145/3328485.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proc. ACM MM*. https://doi.org/10.1145/3503161.3547754.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In *Proc. ACL*. https://doi.org/10.18653/v1/2023.acl-long.421.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proc. AAAI*. https://doi.org/10.1609/aaai.v35i12.17289.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proc. ACL*. https://doi.org/10.18653/v1/P18-1208.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. 2024. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *Proc. ICLR*. https://openreview.net/forum?id=nY9nITZQjc.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. MIntRec: A new dataset for multimodal intent recognition. In *Proc. ACM MM*. https://doi.org/10.1145/3503161.3547906.

Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. 2025. Multimodal classification and out-of-distribution detection for multimodal intent understanding. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2025.3618541.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proc. EMNLP*. https://doi.org/10.18653/v1/2023.emnlp-main.49.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proc. AAAI*. https://ojs.aaai.org/index.php/AAAI/article/view/29656.

Qianrui Zhou, Hua Xu, Yifan Wang, Xinzhi Dong, and Hanlei Zhang. 2025. LLM-Guided semantic relational reasoning for multimodal intent recognition. In *Proc. EMNLP*. https://doi.org/10.18653/v1/2025.emnlp-main.1130.

A Implementation and Dataset Details

Backbones and projection. For the text backbone we use BERT-large-uncased. For audio we use WavLM-base with 768-d frame features mean-pooled over the utterance, and for video Swin-Transformer ROI features with 256-d. All three modalities are projected to a common 512-d space before fusion. The fusion backend is concatenation followed by a two-layer MLP.

Optimisation. Both phases use AdamW for 20 epochs at batch size 64. The OOD experiments follow the `w/ OOS` and `w/o OOS` protocols described in Section 4.3.

Datasets. MIntRec2.0 (Zhang et al., 2024) provides 1,245 dialogues with 15,040 utterances annotated to 30 fine-grained intent classes, split 7:1:1, with approximately 9,300 in-scope and 5,700 out-of-scope utterances; we treat each utterance as single-turn for benchmarking. IEMOCAP-DA extends IEMOCAP (Busso et al., 2008) with 9,416 utterances labelled to 12 emotion-aware dialogue acts; we follow the released train/dev/test split. CMU-MOSI provides 2,199 YouTube monologues split 1,284/229/686, and CMU-MOSEI (Zadeh et al., 2018) provides 23,453 clips in the standard MMSA-FET unaligned protocol (Mao et al., 2022) with the official train/dev/test partition.

Hardware. All experiments were run on a single NVIDIA RTX 5090 GPU.

B Orthogonal Residual Decomposition: Derivations

This appendix derives the two ingredients of Section 3.3 from first principles: the Dirichlet-cross-entropy plus KL supervision used by Residual Geometry Structuring (RGS, §3.2), and the orthogonal decomposition $c = \alpha t + \gamma$ that gives rise to Hierarchical Modality Calibration (HMC) and Asymmetric Multimodal Disentanglement (AMD).

B.1 Dirichlet Cross-Entropy from the Expected Log-Likelihood

Phase 1 treats the text prediction as a distribution over the class-probability simplex rather than as a point estimate. Given evidence $e \in \mathbb{R}^K_{\geq 0}$ and concentration $\boldsymbol{\alpha}_{\text{dir}} = e + \mathbf{1}$, the simplex posterior is $\boldsymbol{p} \sim \text{Dir}(\boldsymbol{\alpha}_{\text{dir}})$. Taking the expected categorical cross-entropy under this Dirichlet,

$$\mathbb{E}_{\boldsymbol{p}\sim\text{Dir}(\boldsymbol{\alpha}_{\text{dir}})}\Big[-\sum_{k=1}^{K} y_k \log p_k\Big] = \sum_{k=1}^{K} y_k\big(\psi(S) - \psi(\alpha_{\text{dir},k})\big), \tag{19}$$

where $S = \sum_k \alpha_{\text{dir},k}$ and we use the standard identity $\mathbb{E}_{\text{Dir}(\boldsymbol{\alpha})}[\log p_k] = \psi(\alpha_k) - \psi(S)$. Equation 19 is exactly $\mathcal{L}_{\text{DCE}}$; minimising it increases target-class evidence relative to off-target evidence. The KL term below separately suppresses unsupported off-target evidence.

The KL term flattens off-target evidence. With the flattened parameters $\tilde{\boldsymbol{\alpha}}_{\text{dir}} = y + (1-y) \odot \boldsymbol{\alpha}_{\text{dir}}$, the KL against a uniform Dirichlet has the closed

Methods	CMU-MOSI					CMU-MOSEI				
	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑
MulT	0.871	0.698	40.00	—/83.0	—/82.8	0.559	0.733	52.84	81.15/84.63	81.56/84.52
MISA	0.783	0.776	42.30	81.80/83.40	81.70/83.60	0.555	0.756	52.20	83.60/85.50	83.80/85.30
Self-MM	0.708	0.796	46.67	83.44/85.46	83.36/85.43	0.531	0.765	53.87	83.76/85.15	83.82/84.90
MMIM	0.700	0.800	46.65	84.14/86.06	84.00/85.98	0.526	0.772	54.24	82.24/85.97	82.66/85.94
FDMER	0.724	0.788	44.10	—/84.60	—/84.70	0.536	0.773	54.10	—/86.19	—/85.80
ConFEDE	0.742	0.784	42.27	84.17/85.52	84.13/85.52	**0.522**	**0.780**	54.86	81.65/85.82	82.17/85.83
ALMT	0.683	0.805	49.42	84.55/86.43	84.57/**86.47**	0.526	0.779	54.28	84.78/**86.79**	85.19/**86.86**
HyDiscGAN	0.749	0.782	43.20	84.10/**86.70**	83.70/86.30	0.533	0.761	54.40	81.90/86.30	82.10/86.20
DEVA	0.737	0.787	46.32	84.40/86.29	84.48/86.31	0.541	0.769	52.26	83.26/86.13	82.93/86.21
AML (Ours)	**0.681**	**0.807**	**49.85**	82.74/84.82	82.60/84.75	**0.522**	0.777	**54.96**	83.04/86.29	82.63/86.23

Table 4: Sentiment-analysis comparison on CMU-MOSI and CMU-MOSEI in the unaligned MMSA-FET protocol. Acc-2 / F1 dual values follow the standard MMSA has-zero / non-zero convention. **Bold** = column best, underline = column second-best.

form

$$\mathrm{KL}\big(\mathrm{Dir}(\tilde{\boldsymbol{\alpha}}_{\mathrm{dir}}) \,\|\, \mathrm{Dir}(\mathbf{1})\big) = \log\Gamma(\tilde{S}) - \log\Gamma(K) - \sum_k \log\Gamma(\tilde{\alpha}_{\mathrm{dir},k}) + \sum_k (\tilde{\alpha}_{\mathrm{dir},k} - 1)\big(\psi(\tilde{\alpha}_{\mathrm{dir},k}) - \psi(\tilde{S})\big), \tag{20}$$

with $\tilde{S} = \sum_k \tilde{\alpha}_{\mathrm{dir},k}$. This term is zero iff the off-target concentrations are all unity (no spurious evidence) and grows monotonically with off-target evidence. The combination of Equations 19–20 jointly supervises the shared text representation and its two output heads during Phase 1 together with the classification loss on t.

B.2 Uniqueness of the Orthogonal Decomposition

Fix any non-zero text logit $t \in \mathbb{R}^K$ and any correction $c \in \mathbb{R}^K$. The decomposition

$$c = \alpha t + \gamma, \qquad \langle \gamma, t \rangle = 0, \tag{21}$$

admits a unique solution. Taking the inner product of Equation 21 with t gives $\langle c, t \rangle = \alpha \langle t, t \rangle$, hence $\alpha = \langle c, t \rangle / \langle t, t \rangle$ and $\gamma = c - \alpha t$; uniqueness follows because $t/\|t\|$ and the orthogonal complement $\{t\}^\perp$ span $\mathbb{R}^K$ as a direct sum. Both quantities are obtained at inference by two inner products and a subtraction, with no additional learnable parameters in HMC—a property that the equation count of Section 3.3 requires but does not derive.

B.3 γ as the Sole Argmax-Shifting Channel

Substituting $c = \alpha t + \gamma$ into the final logits and grouping the t-aligned terms gives

$$\mathrm{final}(x) = \underbrace{\big[\beta + (1-\beta)\alpha\big]}_{\rho}\, t \;+\; (1-\beta)\,\gamma, \tag{22}$$

with mixing $\beta(x) \in (0, 1)$. When $\alpha > 0$ the scalar $\rho = \beta + (1-\beta)\alpha$ is strictly positive, and $\rho\, t$ preserves the argmax of t because positive rescaling preserves the coordinate ordering of a logit vector. Therefore every change in the predicted class $\hat{y} = \arg\max_k \mathrm{final}(x)_k$ relative to the text head's prediction $\hat{y}_{\mathrm{T}} = \arg\max_k t_k$ is induced exclusively by $(1-\beta)\gamma$. Specifically, for any non-target class $j \neq \hat{y}_{\mathrm{T}}$, class j overtakes the text-head winner iff

$$\gamma_j - \gamma_{\hat{y}_{\mathrm{T}}} \;>\; \tfrac{\rho}{1-\beta}\big(t_{\hat{y}_{\mathrm{T}}} - t_j\big). \tag{23}$$

The right-hand side is the text head's confidence margin scaled by $\rho/(1-\beta)$; the residual γ must exceed this margin in the same coordinate direction for the multimodal prediction to diverge from the text. HMC therefore plays a gating role (controlling ρ and hence the margin to overcome), and AMD plays a steering role (deciding through γ where the prediction moves). A second consequence is that Equation 23 is asymmetric across classes when t is not flat: easier flips correspond to lower text margins. The inverse-entropy factor in Equation 16 gives greater regularisation weight to samples for which the text prediction is more certain.

C Generalization to Multimodal Sentiment Analysis

We validate the asymmetric-fusion thesis beyond intent recognition on the multimodal sentiment benchmarks CMU-MOSI and CMU-MOSEI (Zadeh et al., 2018) in the unaligned variant of the MMSA-FET protocol (Mao et al., 2022). For text we use BERT-base-uncased, audio uses COVAREP features, and visual uses Facet features, all from the standard MMSA-FET `unaligned_50.pkl` split. We follow the standard MMSA-FET regression protocol and optimise a smooth-L_1 objective. Standard MMSA metrics are reported: MAE (lower is

better), Pearson correlation, 7-way accuracy (Acc-7), and 2-way accuracy / F1 in the has-zero (h0) and non-zero (n0) conventions.

We compare against nine published baselines: MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021a), FDMER (Yang et al., 2022), ConFEDE (Yang et al., 2023), ALMT (Zhang et al., 2023), HyDiscGAN (Wu et al., 2024), and DEVA (Wu et al., 2025). Numbers for MulT, MISA, Self-MM, ConFEDE, ALMT, and HyDiscGAN are from MSAmba (He et al., 2025) Table 1, which provides MMSA-FET-protocol-standardised reproductions; MMIM is from the official MMSA-FET result repository; FDMER and DEVA are from the original papers.

Results are reported in Table 4. AML achieves the best result on four metrics—MOSI MAE, MOSI Corr, MOSI Acc-7, and MOSEI Acc-7—and matches the best on MOSEI MAE. The MOSI Acc-2 / F1 metrics sit at the third tier of the table; this is a structural ceiling traced to the small MOSI test set (686 samples, single-sample noise approximately 0.15pp), the small MOSI dev set (229 samples, making the per-seed boundary-threshold τ too noisy to transfer), and the low fraction of neutral samples in MOSI (4.4%). The same factors do not apply to MOSEI. The MOSI / MOSEI results support the claim that the asymmetric-fusion design is not narrowly tailored to intent recognition but generalises to other text-dominated multimodal tasks.