REASON, DECOMPOSE, AND DELIBERATE: PARTIAL-INFORMATION-GUIDED MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Sentiment Analysis (MSA) is challenging because textual, acoustic, and visual cues relate to sentiment differently across instances. They may jointly convey a meaning absent from any single modality (e.g., sarcasm), or leave the decisive cue to one modality. To handle this variability, we propose the Reason–Decompose–Deliberate (RDD) framework, which organizes multimodal evidence into three cross-modal roles of redundancy, synergy, and uniqueness. This role-level view turns fusion into an instance-specific decision about how evidence should be combined, so that shared, interaction-dependent, and modality-specific cues are modeled separately rather than entangled in a single representation. To *Reason*, Multimodal Information-Role Reasoning and Routing (MIRR) prompts an MLLM with chain-of-thought to estimate an instance-specific cross-modal role vector and maps it to routing gates. To *Decompose*, Information-Role Guided Expert Decomposition (IGED) extracts redundant, synergistic, and unique features with dedicated experts and fuses them via gate-weighted summation across modalities. To *Deliberate*, Uncertainty-Aware Latent Deliberation (UALD) identifies uncertain samples by predictive entropy and iteratively refines their representations with a latent think token before prediction. Consequently, RDD unifies role reasoning, role-specific decomposition, and uncertainty-driven refinement into a single process that adapts to the evidence configuration of each instance. Experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS show that RDD achieves the best results on all reported metrics.

1 INTRODUCTION

Multimodal Sentiment Analysis (MSA) (Poria et al., 2018) infers human affect by jointly modeling textual, acoustic, and visual cues. While multimodal signals provide complementary emotional evidence, they differ in distributions, temporal dynamics, and semantic expressions. Moreover, their contributions are instance-dependent, ranging from shared and interaction-dependent information to modality-specific cues, making effective multimodal modeling challenging.

Existing MSA research has progressively evolved from feature-level fusion toward explicit cross-modal interaction and structured representation learning. Early methods integrate modalities through tensor fusion (Zadeh et al., 2017), low-rank approximation (Liu et al., 2018), memory networks (Zadeh et al., 2018a), and cross-modal attention (Tsai et al., 2019a). More recent approaches improve multimodal representations through modality-invariant learning (Hazarika et al., 2020), self-supervision (Yu et al., 2021), feature decomposition (Yang et al., 2023a), graph-based interaction (Zhuang et al., 2024), and mutual-information modeling (Sun et al., 2025). Methods (Jian et al., 2026; Dan et al., 2026) further separate shared and modality-specific representations. Despite these advances, most approaches employ fixed interaction or decomposition patterns and lack explicit instance-level reasoning about how multimodal information contributes to each prediction.

This limitation becomes particularly important under mixed multimodal evidence (Fig. 1, top). For instance, aligned positive language, prosody, and facial expressions provide largely redundant sentiment cues, whereas sarcasm may arise from interactions between positive text and contradictory acoustic or visual signals. In other cases, sentiment may depend primarily on a unique cue from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

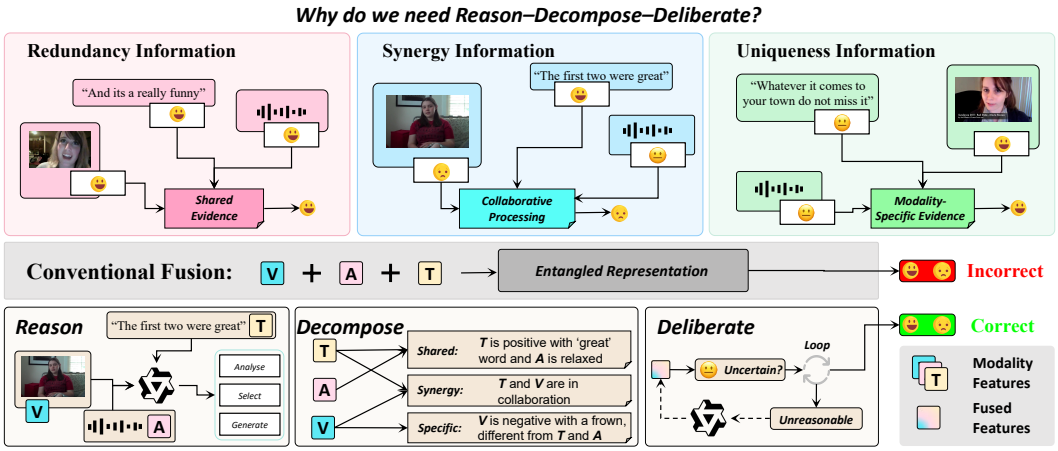


Figure 1: Motivation of RDD. Multimodal evidence may be redundant, synergistic, or modality-specific, making uniform fusion inadequate. RDD reasons about these roles, decomposes them through specialized experts, and deliberates over uncertain representations.

a single modality. Treating such distinct relationships within a unified fusion space can entangle shared, interaction-dependent, and modality-specific information. Moreover, most MSA models still follow a single-pass inference paradigm: once multimodal features are fused, the prediction is produced directly, leaving little opportunity to reconsider uncertain interpretations.

Recent advances in large-model reasoning provide a complementary perspective. Chain-of-Thought (CoT) prompting (Wei et al., 2022) and its extensions toward latent computation and iterative refinement (Goyal et al., 2024; Geiping et al., 2026; Li et al., 2026; Du et al., 2026; Ye et al., 2026) suggest two underexplored opportunities for MSA: using MLLMs to reason about cross-modal information relationships rather than directly predicting sentiment, and allowing uncertain representations to undergo latent deliberation before prediction.

Motivated by these observations, we propose the Reason–Decompose–Deliberate (RDD) framework (Fig. 1, bottom) for MSA. Inspired by Partial Information Decomposition (PID) (Williams & Beer, 2010), RDD organizes multimodal affective information into three cross-modal roles: redundancy, synergy, and uniqueness, corresponding to shared, interaction-dependent, and modality-specific cues. RDD first reasons about the dominant information roles of each instance, then decomposes the representation accordingly, and finally deliberates over uncertain cases. Specifically, Multimodal Information-role Reasoning and Routing (MIRR) employs an MLLM with CoT prompting to infer instance-specific role priors and translate them into adaptive routing signals. Information-role Guided Expert Decomposition (IGED) then routes representations through specialized Redundancy, Synergy, and Uniqueness experts for structured disentanglement. Finally, Uncertainty-Aware Latent Deliberation (UALD) identifies ambiguous samples via predictive uncertainty and iteratively refines their latent states, with a feedback mechanism re-evaluating high-confidence failures to update role priors. Extensive experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS consistently validate the effectiveness of RDD against recent methods.

In summary, our main contributions are as follows:

- We propose the Reason–Decompose–Deliberate (RDD) framework, which unifies cross-modal role reasoning, structured decomposition, and uncertainty-aware deliberation for multimodal sentiment analysis.
- We introduce MIRR and IGED to leverage MLLM-derived cross-modal role vectors for adaptively routing representations into Redundancy, Synergy, and Uniqueness experts, enabling interpretable information decomposition.
- We develop UALD with a feedback mechanism to iteratively refine uncertain representations and correct unreliable role reasoning through closed-loop optimization.
- Experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS demonstrate that RDD consistently achieves the strongest performance across all reported metrics.

2 Related Work

2.1 Multimodal Sentiment Analysis (MSA)

MSA has progressed from feature-level fusion to explicit cross-modal interaction modeling. Early works adopted tensor fusion (Zadeh et al., 2017), low-rank approximation (Liu et al., 2018), and dynamic weighting (Du et al., 2025), followed by memory networks (Zadeh et al., 2018a), cross-modal attention (Tsai et al., 2019a), directional interaction (Wang et al., 2022), tensor extensions (Wang et al., 2023), and global-local fusion graphs (Zhuang et al., 2024). (Li et al., 2025) further estimates cross-modal dependencies before fusion. Yet these methods strengthen interaction modeling without distinguishing different information roles across modalities.

Recent work turns to disentanglement and complementary information modeling. (Hazarika et al., 2020) separates modality-invariant and modality-specific representations, while (Yu et al., 2021) improve consistency via self-supervision and federated learning. (Jian et al., 2026) and (Dan et al., 2026) further decompose shared and modality-specific signals, and (Sun et al., 2025) adopts mutual-information objectives to characterize invariant, specific, and complementary components. Other works enhance robustness through consistency (Yu et al., 2020), efficient interaction (Sun et al., 2023), graph reasoning (Mai et al., 2023), feature enrichment (Wu et al., 2025), masked multi-view learning (Wang et al., 2025), and external knowledge (Zhu et al., 2024) (Zhang et al., 2025). Yet most rely on fixed decompositions and lack instance-level reasoning over redundant, synergistic, and unique information, motivating our RDD framework.

2.2 Chain-of-Thought (CoT)

CoT prompting elicits intermediate reasoning steps before final prediction (Wei et al., 2022). Follow-up work explains it theoretically (Feng et al., 2023) and transfers it to smaller models via reasoning distillation (Ho et al., 2023) and symbolic distillation (Li et al., 2023b). The reasoning process is further optimized via prompt rewriting (Chen et al., 2024), self-composed structures (Zhou et al., 2024), and tree-path preference optimization (Zhang et al., 2024), while (Wang & Zhou, 2024) surfaces latent CoT paths without manual prompts. Efficiency is pursued through compressed traces (Kang et al., 2025) and adaptive budgets (Wu et al., 2026). These motivate our MIRR module, which uses structured CoT to infer instance-specific redundancy, synergy, and uniqueness priors.

A complementary direction internalizes reasoning into continuous latent states. (Yao et al., 2023) introduces explicit multi-path search, while pause tokens (Goyal et al., 2024) and recurrent-depth computation (Geiping et al., 2026) enable hidden deliberation without long verbal traces. Recent work adapts reasoning depth (Li et al., 2026), optimizes latent trajectories directly (Du et al., 2026) (Ye et al., 2026), and extends to multimodal latent reasoning (Ray et al., 2026) and self-correcting vision-language models (Ding & Zhang, 2026). Inspired by this line, our UALD module performs uncertainty-triggered latent deliberation on ambiguous samples.

2.3 Multimodal Fusion

Multimodal fusion has progressed from tensor interactions to pretrained alignment. Early multiplicative fusion suffered from exponential dimensionality, addressed by (Liu et al., 2018) through low-rank factors, while (Tsai et al., 2019b) replaced static tensors with directional cross-modal attention for unaligned streams. Pretraining further advanced visual-linguistic fusion: (Lu et al., 2019) introduced two-stream co-attention, (Tan & Bansal, 2019) separated language, object, and cross-modality encoders, (Li et al., 2020) anchored region-text alignment, and (Dou et al., 2022) incorporated cross-attention into backbones.

With the rise of large multimodal models, fusion focuses on efficiently bridging frozen encoders. (Alayrac et al., 2022) injects visual features into frozen language models via gated cross-attention, (Li et al., 2023a) uses a lightweight Querying Transformer for modality alignment, and (Girdhar et al., 2023) aligns six modalities in a unified space. However, these models assume complete modalities and degrade under missing inputs (Ma et al., 2022), motivating methods that exploit prompt correlations (Hu et al., 2024) or specialized subnetworks (Shen et al., 2023). Existing fusion methods mainly capture global interactions without separating shared, emergent, and modality-specific information. Since cross-modal relations vary by instance, role-aware adaptive fusion is needed.

3 METHODOLOGY

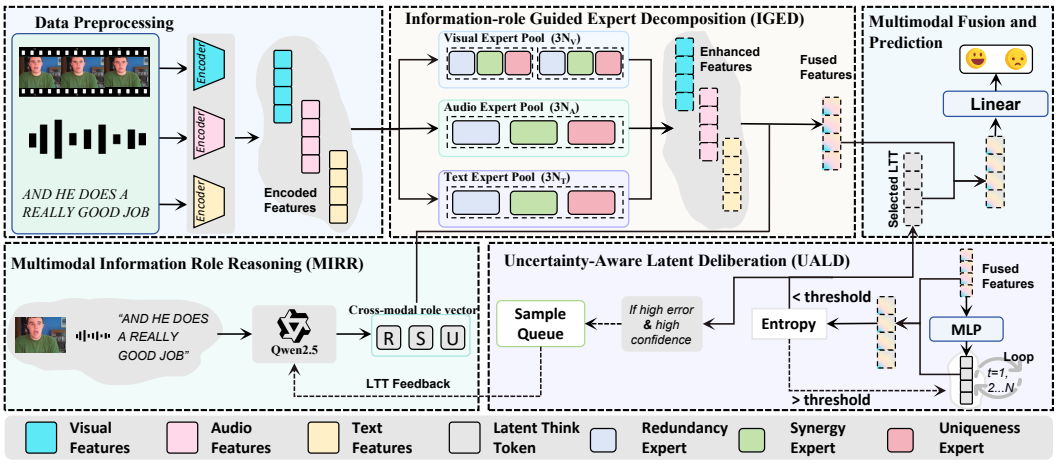


Figure 2: Overview of the proposed RDD framework. MIRR *reasons* about instance-specific cross-modal roles, IGED *decomposes* multimodal representations through specialized experts, and UALD *deliberates* over uncertain cases through iterative latent refinement and feedback.

Fig. 2 presents the proposed Reason–Decompose–Deliberate (RDD) framework, which models multimodal information using five components: data preprocessing, MIRR, IGED, UALD, and multimodal fusion and prediction. In data preprocessing, text is encoded with pre-trained BERT (Devlin et al., 2019), while acoustic and visual features are aligned to the text sequence. 1D convolutions and modality-specific Transformer encoders produce x_m, $m \in \{T, A, V\}$. Next, MIRR infers instance-specific role priors, IGED performs role-guided decomposition, and UALD refines uncertain representations. Finally, multimodal fusion and prediction integrates refined representations and output the final sentiment prediction.

3.1 MULTIMODAL INFORMATION-ROLE REASONING AND ROUTING (MIRR)

To capture instance-specific multimodal interactions beyond static fusion, MIRR employs a MLLM (Jin Xu, 2025) with CoT prompting (Wei et al., 2022) to infer cross-modal role vectors over redundancy, synergy, and uniqueness. These priors are subsequently converted into continuous routing signals for downstream expert decomposition.

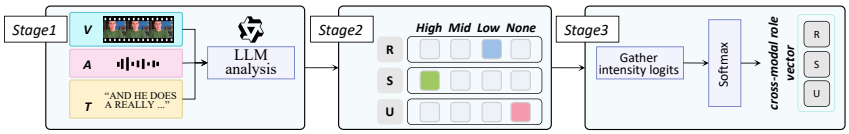


Figure 3: Pipeline of MIRR. MLLM reasoning over Redundancy ($\mathcal{R}$), Synergy ($\mathcal{S}$), and Uniqueness ($\mathcal{U}$) is converted into a continuous 3D prior through logit-level extraction.

As shown in Fig. 3, we adopt a three-stage CoT procedure. **Stage 1:** The MLLM jointly analyzes the complete multimodal input while considering all modalities equally, preventing dominant modalities from biasing the estimation of cross-modal information roles. **Stage 2:** It evaluates Redundancy ($\mathcal{R}$), Synergy ($\mathcal{S}$), and Uniqueness ($\mathcal{U}$) using four intensity levels: *high*, *medium*, *low*, and *none*, associated with $\mathbf{w} = [1, \frac{2}{3}, \frac{1}{3}, 0]^\top$. These roles capture key multimodal patterns: $\mathcal{R}$ denotes shared evidence, $\mathcal{S}$ reflects interaction-dependent information, and $\mathcal{U}$ represents modality-specific cues. Rather than estimating the full PID lattice, we adopt them as an interpretable abstraction. The four ordinal levels provide a reliable semantic scale for MLLM judgment and are converted into continuous routing priors. **Stage 3:** The expected intensity $\tilde{p}_o$ of each state $o \in \{\mathcal{R}, \mathcal{S}, \mathcal{U}\}$ is computed as $\tilde{p}_o = \mathbf{w}^\top \boldsymbol{\pi}_o$, where $\boldsymbol{\pi}_o \in \mathbb{R}^4$ denotes the softmax-normalized probability vector over the four intensity levels derived from the MLLM's output logits for state o. In addition to the role scores, MLLM outputs a self-reported confidence score $r_{\text{LLM}} \in [0, 1]$ that reflects its certainty in the cross-modal role assignment, which is used to weight the reward signal and trigger offline feedback.

The three expected intensities are then normalized to obtain the cross-modal role vector:

$$\mathbf{p} = \frac{[\tilde{p}_{\mathcal{R}}, \tilde{p}_{\mathcal{S}}, \tilde{p}_{\mathcal{U}}]^{\top}}{\sum_{o \in \{\mathcal{R},\mathcal{S},\mathcal{U}\}} \tilde{p}_o}. \tag{1}$$

where $\mathbf{p}$ denotes the importance of redundancy, synergy, and uniqueness for each instance.

Finally, a lightweight routing network maps $\mathbf{p}$ into modality-specific gating scores using linear projection, GELU activation, normalization, and gating heads. These scores form $\boldsymbol{\beta} = [\boldsymbol{\beta}_T; \boldsymbol{\beta}_V; \boldsymbol{\beta}_A]$, which adaptively weights the corresponding expert outputs and transfers MLLM-derived information-role reasoning to subsequent expert decomposition in Sec. 3.2.4. The motivation of the PID-inspired cross-modal roles is provided in Sec. A.1.

3.2 Information-Role Guided Expert Decomposition (IGED)

Standard fusion often entangles shared, synergistic, and modality-specific cues, yielding noise-sensitive representations. To separate these roles, we construct a PID-inspired Orthogonal Operator Space (Bansal et al., 2018) with dedicated Redundancy, Synergy, and Uniqueness experts for each modality, whose outputs are adaptively aggregated into a purified representation. Further interpretation of how these experts realize the three cross-modal roles is provided in Appendix Sec. A.2.

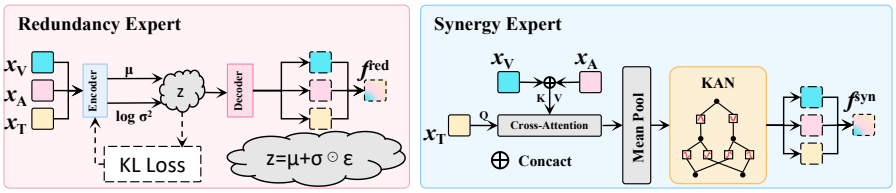


Figure 4: Architectures of the $\mathcal{R}$-expert (left) and $\mathcal{S}$-expert (right), which respectively extract shared multimodal cues and model cross-modal interactions.

3.2.1 Redundancy Expert ($\mathcal{R}$-expert)

The $\mathcal{R}$-expert (Fig. 4 left) is designed to capture affective cues shared across modalities while suppressing modality-specific variations. For each modality $m \in \{T, V, A\}$, the encoder maps x_m to a Gaussian posterior parameterized by mean μ_m and standard deviation σ_m. A stochastic latent representation z_m is then sampled using the reparameterization trick:

$$z_m = \mu_m + \sigma_m \odot \epsilon_m, \qquad \epsilon_m \sim \mathcal{N}(0, I). \tag{2}$$

where $\odot$ denotes element-wise multiplication and ϵ_m is standard Gaussian noise.

A deterministic decoder subsequently transforms z_m into the modality-specific redundancy representation f_m^{red}. To regularize the latent space, the posterior is constrained toward a standard normal distribution using the Kullback-Leibler (KL) (Kingma & Welling, 2013) divergence:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum_i \left(1 + \log \sigma_i^2 - \mu_i^2 - \sigma_i^2\right), \tag{3}$$

where μ_i and σ_i denote the mean and standard deviation of the i-th latent dimension. $\mathcal{L}_{\text{KL}}$ is summed over latent dimensions per sample and averaged across the mini-batch during training.

A mini-batch contrastive loss $\mathcal{L}_{\text{align}}$ aligns redundancy latents by treating cross-modal representations from the same sample as positives and others as negatives:

$$\mathcal{L}_{\text{align}} = -\frac{1}{3} \sum_{(m,m') \in \{(T,V),(T,A),(V,A)\}} \log \frac{\exp(\text{sim}(z_m, z_{m'}))}{\sum_z \exp(\text{sim}(z_m, z))}, \tag{4}$$

where z_m denotes the redundancy latent of modality $m \in \{T, V, A\}$, $z_{m'}$ is its positive counterpart from the same sample, z denotes each redundancy latent in the batch over which the denominator sum is taken (including both the positive and all negative samples), and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. This encourages compact and consistent shared affective information across modalities.

3.2.2 Synergy Expert ($\mathcal{S}$-expert)

While $\mathcal{R}$-expert captures consensus, higher-order semantics (e.g., sarcasm or irony) often emerge exclusively from non-linear, cross-modal collisions. The $\mathcal{S}$-expert (Fig. 4 right) is specifically designed to model these complex emergent interactions.

First, text features serve as queries to attend to concatenated visual and acoustic features, yielding $h_{\text{cross}} = \text{Softmax}\left(\frac{QW_Q(KW_K)^\top}{\sqrt{d}}\right)KW_V$, where Q, K, and V denote the query, key, and value matrices projected by W_Q, W_K, and W_V, respectively, and $\sqrt{d}$ scales the attention logits.

The resulting cross-modal interaction features are pooled into h_{pool}, which summarizes the most informative interaction patterns across modalities. A KAN-inspired nonlinear basis projector then transforms h_{pool} into the synergistic representation f^{syn}:

$$f^{\text{syn}} = \Phi^{(L)}(\Phi^{(L-1)}(\cdots(\Phi^{(1)}(h_{\text{pool}}))), \qquad [\Phi^{(l)}(x)]_j = \sum_i \phi_{ji}^{(l)}(x_i), \tag{5}$$

where $\phi_{ji}^{(l)}(\cdot)$ denotes a learnable nonlinear basis function that maps the i-th input to the j-th output at layer l. Compared with a fixed linear projection, this composition more flexibly models nonlinear dependencies in cross-conditioned features, making it suitable for synergy arising from joint multimodal interactions. More implementation details are provided in the Appendix Sec. A.3.

3.2.3 Uniqueness Expert ($\mathcal{U}$-expert)

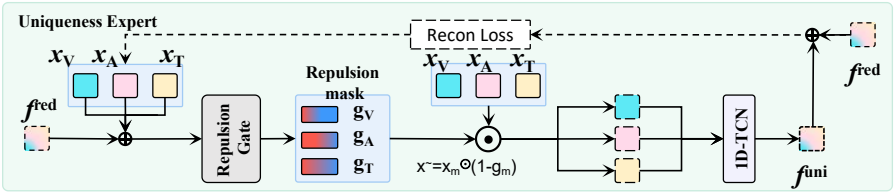


Figure 5: $\mathcal{U}$-expert. A redundancy-conditioned gate suppresses overlap, while a 1D-TCN extracts modality-specific cues and reconstruction preserves fidelity.

Beyond shared and interacting signals, individual modalities often contain exclusive, decisive cues. The $\mathcal{U}$-expert (Fig. 5) isolates this modality-specific information while actively suppressing signals that have already been explained by the $\mathcal{R}$-expert.

First, given the raw sequence x_m and redundancy f_m^{red}, a dynamic repulsion gate concatenates both as input and produces a mask $g_m = Sigmoid\big(W_g[x_m; f_m^{\text{red}}] + b_g\big)$, where W_g and b_g are learnable parameters. The raw feature is then purified as $\widetilde{x}_m = x_m \odot (1 - g_m)$, which suppresses redundant components that overlap with f_m^{red}.

A 1D-TCN (Bai et al., 2018) processes $\widetilde{x}_m$ to capture modality-specific dynamics, yielding f_m^{uni}. To preserve critical information, an auxiliary reconstruction branch recovers original representation as $\widehat{x}_m = \text{Dec}_{\text{rec}}\left([f_m^{\text{uni}}; f_m^{\text{red}}]\right)$, where Dec_{rec} denotes the reconstruction decoder.

Finally, the reconstruction objective $\mathcal{L}_{\text{recon}} = \|\widehat{x}_m - x_{\text{m}}\|_2^2$ prevents trivial or noisy solutions by requiring $[f_m^{\text{uni}}; f_m^{\text{red}}]$ to reconstruct the unperturbed input x_{m}, encouraging f_m^{uni} to retain discriminative cues complementary to the redundancy representation.

3.2.4 Information-Role Guided Expert Aggregation

Within IGED, $\mathcal{R}$-expert, $\mathcal{S}$-expert, and $\mathcal{U}$-expert form modality-specific triplets. To reflect different modality capacities, visual, audio, and textual pools contain $3N_V$, $3N_A$, and $3N_T$ experts, respectively. For each modality m, the router adaptively weights the three information-role representations f_m^{red}, f^{syn}, and f_m^{uni}, which are aggregated across modalities as $f^{\text{fused}} = \sum_{m\in\{T,V,A\}}\left(\beta_{\text{red}}^m f_m^{\text{red}} + \beta_{\text{syn}} f^{\text{syn}} + \beta_{\text{uni}}^m f_m^{\text{uni}}\right)$. The resulting f^{fused} adaptively integrates redundant, synergistic, and unique affective information while suppressing irrelevant noise, providing a structured multimodal representation for subsequent latent deliberation.

3.3 Uncertainty-Aware Latent Deliberation (UALD)

Single-pass fusion may be unreliable for ambiguous or conflicting multimodal samples. We therefore introduce a differentiable refinement mechanism for uncertain representations.

At step t, the fused representation f^{fused} is augmented with the current Latent Think Token (LTT) state f_t^{LTT}, yielding the residual representation $h_t = f^{\text{fused}} + f_t^{\text{LTT}}$, which is fed to the sentiment

predictor. Predictive uncertainty is quantified by the entropy of the resulting prediction, $\mathcal{H}(h_t)$, where $\mathcal{H}(\cdot)$ denotes the entropy function. If $\mathcal{H}(h_t) < \gamma$, the refinement stops. Otherwise, LTT is updated toward more confident candidates:

$$\epsilon^b \sim \mathcal{N}(0, \sigma^2 I), \qquad w_t^b = \frac{\alpha_t^b}{\sum_{j=1}^{B} \alpha_t^j}, \qquad f_{t+1}^{\mathrm{LTT}} = f_t^{\mathrm{LTT}} + \lambda \sum_{b=1}^{B} w_t^b \epsilon^b, \tag{6}$$

where γ is the uncertainty threshold, σ controls the exploration scale, and I is the identity matrix. α_t^b denotes the confidence of candidate b, w_t^b its normalized weight, and λ the update step size.

The above process is repeated for at most N steps or until early termination, after which the selected LTT f^{LTT} is used to refine the fused representation for final prediction, which acts as an adaptive residual that refines f^{fused} toward more confident latent states.

To mitigate confidently wrong predictions, a feedback mechanism flags samples with large prediction error, measured as the absolute difference between the ground-truth sentiment label and the network prediction, but high MLLM confidence r_{LLM}, indicating potentially incorrect cross-modal role assignments. For each flagged sample, a corrective prompt provides the sample id, ground-truth sentiment, and network prediction, asking the MLLM to reconsider overlooked synergy (e.g., irony or sarcasm) or overestimated modality-unique cues in MIRR. These prompts are accumulated during training and re-evaluated offline, updating the cached Redundancy/Synergy/Uniqueness role vectors to reduce similar routing errors in subsequent epochs and form a closed-loop refinement process. The interpretation as stochastic latent optimization is provided in Appendix Sec. A.4.

3.4 Multimodal Fusion and Prediction

The selected LTT f^{LTT} provides an adaptive residual correction to f^{fused}, yielding the final representation f^{final}. Thus, the expert decomposition constructs the primary multimodal representation, while UALD further refines difficult cross-modal cases before prediction.

The overall objective combines the supervised task loss with auxiliary expert regularization as $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{recon}}\mathcal{L}_{\mathrm{recon}}$. Here, $\mathcal{L}_{\mathrm{task}}$ supervises sentiment prediction, while the auxiliary losses promote information-role separation and reconstruction fidelity.

4 Experiments

4.1 Datasets and Implementation Details

Our method is evaluated on CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018b), and CH-SIMS (Yu et al., 2020). More details are provided in Appendix A.10. All experiments are conducted on a single NVIDIA A6000 GPU with the Adam optimizer (Kingma & Ba, 2014). We set batch size=16, learning rate=2.8×10^{-5}, weight decay=7×10^{-7}, and maximum epochs=100. The expert pool sizes are N_T=2, N_V=3, N_A=2. The KAN projector uses L=2 and G=5. For UALD, $\lambda = 10^{-3}$, σ=0.1, B=3, N=5, and γ=0.3. And λ_{align}=0.05, λ_{KL}=0.01, λ_{recon}=0.05. The feedback thresholds are prediction error $= 1.2$ and r_{LLM}=0.8.

4.2 Comparison with the State-of-the-art Methods

As presented in Table 1, our framework achieves exceptional performance across the majority of evaluation metrics on CMU-MOSI and CMU-MOSEI datasets. Specifically, on the MOSI dataset, our model yields an Acc_2 of 87.82% and an F1 score of 87.79%, outperforming the second-best baseline CACR (87.04% Acc_2 and 86.90% F1) by solid margins of 0.78% and 0.89%, respectively. For fine-grained evaluation, our Acc_7 reaches 49.54%, surpassing the second-best Fuse-Net (49.27%) by 0.27%, while our MAE is pushed down to 0.673, superior to the second-best PSA-MF (0.686). Similarly, on the MOSEI dataset, our model secures the top rank in binary metrics with an Acc_2 of 87.23% and an F1 score of 87.25%, outperforming MMSLF (86.62%) and MDF (86.61%). Furthermore, our approach achieves an Acc_7 of 55.74%, surpassing the second-best competitors PSA-MF and GLoMo (both at 55.00%), and attains an MAE of 0.512, outperforming the second-best PSA-MF (0.521). On the CH-SIMS dataset, our model again ranks first across all five metrics as shown in Table 2. It attains an Acc_2 of 82.54%, outperforming the second-best FRSR (81.71%) by 0.83%,

Table 1: Performance comparison on MOSI and MOSEI. Best results are in **bold**, second-best are underlined, and third-best are marked with *.

Method	MOSI					MOSEI				
	$Acc_2 \uparrow$	F1 ↑	$Acc_7 \uparrow$	Corr ↑	MAE ↓	$Acc_2 \uparrow$	F1 ↑	$Acc_7 \uparrow$	Corr ↑	MAE ↓
MISA (Hazarika et al., 2020)	83.40	83.60	42.30	0.761	0.783	85.50	85.30	52.20	0.756	0.555
Self-MM (Yu et al., 2021)	84.90	84.90	45.30	–	0.738	84.50	84.30	53.20	–	0.540
EMT (Sun et al., 2023)	85.00	85.00	47.40	0.798	0.705	86.00	86.00	54.50	0.774	0.527
ConFEDE (Yang et al., 2023b)	85.50	85.50	42.30	0.782	0.742	85.80	85.80	54.90*	0.780*	0.522*
GLoMo (Zhuang et al., 2024)	86.70	86.60	48.30	0.782	0.718	86.50	86.40	55.00	0.771	0.539
MFON (Zhang et al., 2025)	86.90*	86.90	44.90	0.797	0.725	86.30	86.30	53.70	0.780*	0.528
DecAlign (Qian et al., 2026)	85.75	85.82	–	–	0.735	86.48	86.07	–	–	0.543
MMSLF (Zhang et al., 2026)	86.61	86.69*	–	0.797	0.734	86.62	86.71	–	0.773	0.539
PSA-MF (Xie et al., 2026)	86.43	86.19	46.50	0.807*	0.686	86.30	86.28	55.00	0.774	0.521
MDF (Jian et al., 2026)	85.82	85.90	48.25	0.810	0.692	86.61*	86.41*	54.80	0.788	0.525
CACR (Wang et al., 2026)	87.04	86.90	48.69	–	0.710	85.37	85.36	54.41	–	0.532
TSD (Meng et al., 2026)	86.50	86.60	49.00*	–	0.697	86.20	86.20	54.60	–	0.525
Fuse-Net (Yang & Li, 2026)	85.83	85.96	49.27	0.798	0.688*	85.88	86.14	54.32	0.772	0.527
RDD(Ours)	**87.82**	**87.79**	**49.54**	**0.825**	**0.673**	**87.23**	**87.25**	**55.74**	**0.793**	**0.512**

while its Acc_3 and Acc_5 reach 68.30% and 45.95%, surpassing the second-best FENet (68.05% and 44.42%) by 0.25% and 1.53%, respectively. The MAE is further pushed down to 0.379 (vs. 0.403 of the second-best DecAlign), and the correlation coefficient reaches 0.683, substantially exceeding the second-best CADET (0.628) by a clear margin.

The consistent gains across benchmarks stem from the RDD design. MIRR provides instance-specific cross-modal role vectors, IGED reduces entanglement through structured expert decomposition, and UALD uses Latent Think Tokens to refine ambiguous multimodal conflicts with feedback on unreliable routing. Together, these modules enable more structured and noise-resistant reasoning, improving prediction accuracy.

Table 2: Performance comparison on CH-SIMS. Best results are in **bold**, second-best are underlined, and third-best are marked with *.

Method	$Acc_2 \uparrow$	$Acc_3 \uparrow$	$Acc_5 \uparrow$	MAE↓	Corr↑
Self-MM (Yu et al., 2021)	80.04	65.47	41.53	0.425	0.595
MMIM (Han et al., 2021)	78.30	65.50	42.00	0.422	0.597
TETFN (Wang et al., 2023)	81.18*	63.24	41.79	0.420	0.577
EMT (Sun et al., 2023)	80.10	67.4*	43.50	0.396	0.623
HCIL (Fu et al., 2024)	79.17	–	–	0.431	0.595
FNENet (Zheng et al., 2024)	80.31	68.05	44.42	0.417	0.618
DEVA (Wu et al., 2025)	79.64	65.42	43.07	0.424	0.583
SSLMM (Wang et al., 2025)	80.56	–	39.52	–	0.589
DecAlign (Qian et al., 2026)	–	–	–	0.403	–
FRSR (Zhu et al., 2026)	81.71	–	–	0.408*	0.626*
CADET (Dan et al., 2026)	80.40	–	44.10*	–	0.628
CACR (Wang et al., 2026)	81.62	66.7	–	0.423	0.596
RDD(Ours)	**82.54**	**68.30**	**45.95**	**0.379**	**0.683**

4.3 Ablation Study

In w/o MIRR, the MLLM-derived semantic prior is replaced by static learnable weights while retaining the $\mathcal{R}$-/$\mathcal{S}$-/$\mathcal{U}$-experts; in w/o UALD, the active inference loop is removed and prediction relies directly on the fused representation. We also independently remove $\mathcal{L}_{KL}$, $\mathcal{L}_{align}$, and $\mathcal{L}_{recon}$ to assess redundancy smoothing, cross-modal alignment, and modality-specific fidelity. Additional analyses about auxiliary-loss sensitivity, internal component ablations, comparisons with conventional alternatives, and expert-pool capacity, are provided in Appendix Sec. A.5–A.9.

As reported in Table 3, every component contributes critically. The model w/o UALD causes the largest drop (Acc_2 down 1.57%, MAE rising to 0.715), confirming that single-pass inference struggles with ambiguous cross-modal cues. The model w/o MIRR reduces Acc_2 by 1.39%, validating the inductive value of structured priors. Among the auxiliary losses, dropping $\mathcal{L}_{recon}$ is most detrimental, confirming its role in preventing Uniqueness-branch collapse; removing $\mathcal{L}_{KL}$ and $\mathcal{L}_{align}$ also degrade performance, indicating that the three regularizers are complementary and jointly needed to keep the experts from re-entangling.

Table 3: Architectural ablations on MOSI.

Setting	Acc2↑	F1↑	Acc7↑	MAE↓	Corr↑
w/o MIRR	86.43	86.46	48.17	0.697	0.773
w/o UALD	86.25	86.17	47.85	0.715	0.770
w/o $\mathcal{L}_{KL}$	86.30	86.28	47.57	0.692	0.775
w/o $\mathcal{L}_{align}$	86.53	86.48	47.88	0.695	0.776
w/o $\mathcal{L}_{recon}$	86.17	86.25	47.79	0.703	0.775
Full Model	**87.82**	**87.79**	**49.54**	**0.673**	**0.825**

4.4 VISUALIZATIONS

4.4.1 T-SNE VISUALIZATION

We project the test-set representations produced by the three cross-modal role experts ($\mathcal{R}$-/$\mathcal{S}$-/$\mathcal{U}$-experts) onto a 2-D t-SNE space for each modality separately, as shown in Fig. 6 (left). For the text, audio, and vision branches (top-left, top-right, and bottom-left panels), the three experts consistently form three well-separated, non-overlapping clusters with compact internal structures, indicating that the orthogonal expert routing decomposes each modality's features along their information-theoretic roles—shared consensus, joint synergy, and modality-exclusive cues.

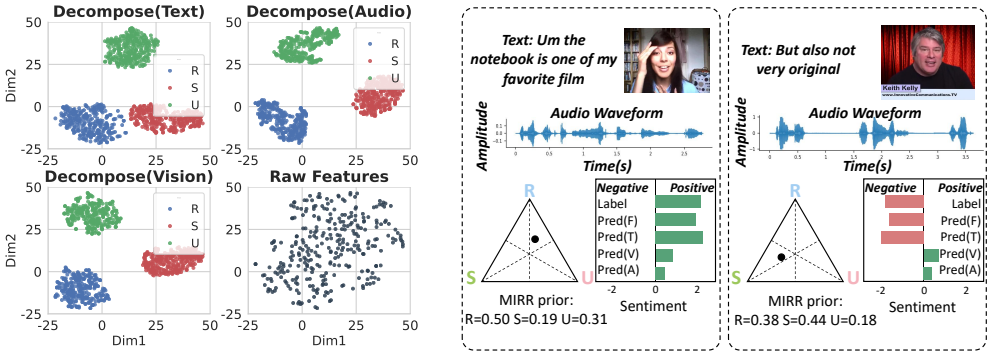


Figure 6: Left: t-SNE (Van der Maaten & Hinton, 2008) visualization of cross-modal role expert representations. Right: two MOSI cases showing MIRR priors and prediction decomposition across redundancy, synergy, and uniqueness.

The bottom-right panel shows the raw encoded features prior to expert-level decomposition, which form a single entangled distribution with no discernible cluster structure. The sharp contrast between the unstructured raw distribution and the three cleanly separated clusters confirms that our cross-modal role routing actively disentangles the feature space along information-theoretic roles, rather than relying on pre-existing geometric separation.

4.4.2 CASE STUDY

Fig. 6 (Right) visualizes the MIRR→IGED reasoning chain on two MOSI samples. Each panel shows the video frame, the MIRR prior as a ternary plot, the comparison between the ground-truth label and the fused/unimodal predictions, and the corresponding audio waveform at the bottom.

In the first sample, all three modalities consistently express positive sentiment (*Pred(T)*=$+2.49$, *Pred(V)*=$+0.61$, *Pred(A)*=$+0.37$); the MIRR prior assigns the highest weight to redundancy ($R = 0.50$), and the fused prediction ($+2.16$) closely matches the label ($+2.40$). In the second sample, the text conveys clear negative sentiment (*Pred(T)*=-2.24) while vision and audio remain mildly positive (*Pred(V)*=$+0.70$, *Pred(A)*=$+0.64$), creating a cross-modal conflict; the prior detects this inter-modal discrepancy and elevates the synergy expert to the highest weight ($S = 0.44$), so that the *inconsistency* across modalities rather than their redundancy drives fusion. Accordingly, the fused prediction (-1.92) resolves the conflict in the negative direction and closely matches the label (-2.00), verifying the accuracy of the MIRR routing.

5 CONCLUSION

We introduced Reason–Decompose–Deliberate (RDD), a modular framework for progressively reasoning about, disentangling, and refining multimodal information. MIRR uses an MLLM to infer instance-specific priors over cross-modal redundancy, synergy, and uniqueness. IGED then routes representations through specialized experts to model shared, interaction-dependent, and modality-specific affective cues. UALD further identifies ambiguous samples via predictive uncertainty and iteratively refines their latent representations, while a feedback mechanism revisits high-confidence routing failures. Experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS demonstrate consistent improvements over existing methods.

AI USE STATEMENT

In this work, we used generative AI tools (LLM-based writing assistants) solely for English language polishing and grammatical refinement of the manuscript text. We have reviewed all AI-assisted writing carefully: every statement, experimental result, claim, and artifact was verified by the authors, and no figure, table, equation, or numerical result was generated or altered by the AI tool. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? *Advances in neural information processing systems*, 31, 2018.

S Chen, B Li, and D Niu. Boosting of thoughts: trial-and-error problem solving with large language models (2024). *arXiv preprint arXiv:2402.11140*, 2024.

Xiangrui Dan, Bin Gao, Yanping Chen, Qian Fu, Yutong Li, Shutian Liu, and Zhengjun Liu. Cadet: Cycle-consistent adversarial disentanglement with text-guided enhancement for multimodal sentiment analysis. *IEEE TAFFC*, 17(3):2917–2930, 2026. doi: 10.1109/TAFFC.2026.3707891.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 4171–4186, 2019.

Yi Ding and Ruqi Zhang. Sherlock: Self-correcting reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 38:101638–101672, 2026.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.

Hanwen Du, Yuxin Dong, and Xia Ning. Latent thinking optimization: Your latent reasoning language model secretly encodes reward signals in its latent thoughts. In *International Conference on Learning Representations*, volume 2026, pp. 101000–101032, 2026.

Quanqi Du, Loic De Langhe, Els Lefever, and Veronique Hoste. Ldw: Label divergence weighting for multimodal sentiment analysis. In *ACM MM*, pp. 12342–12351, 2025.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in neural information processing systems*, 36:70757–70798, 2023.

Yanping Fu, Zhiyuan Zhang, Ruidi Yang, and Cuiyou Yao. Hybrid cross-modal interaction learning. *Neurocomputing*, 571:127201, 2024.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *Advances in Neural Information Processing Systems*, 38: 41340–41391, 2026.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, volume 2024, pp. 27896–27923, 2024.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization. In *EMNLP*, pp. 9180–9192, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM MM*, pp. 1122–1131, 2020.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 14852–14882, 2023.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. *Advances in Neural Information Processing Systems*, 37:67446–67466, 2024.

Zhongquan Jian, Wenhan Lv, Yanhao Chen, Guanran Luo, Wentao Qiu, Shaopan Wang, Bingbing Hu, and Qingqiang Wu. Mdf: A modality-aware disentanglement and fusion framework. In *AAAI*, volume 40, pp. 31292–31300, 2026.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24312–24320, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023a.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2665–2679, 2023b.

Pingzhi Li, Bairu Hou, Yun Zhu, Yihao Feng, Ke Ye, Tao Lei, Zhifeng Chen, Tianlong Chen, and Xianzhi Du. Adaptive thinking: Large language models know when to think in latent space. In *International Conference on Learning Representations*, volume 2026, pp. 93424–93440, 2026.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision*, pp. 121–137. Springer, 2020.

Yangmin Li, Ruiqi Zhu, and Wengen Li. Cormult: A semi-supervised modality correlation-aware multimodal transformer for sentiment analysis. *IEEE TAFFC*, 16(3):2321–2333, 2025. doi: 10.1109/TAFFC.2025.3559866.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, pp. 2247–2256, 2018.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18156–18165. IEEE, 2022.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE TAFFC*, 14(3):2276–2289, 2023. doi: 10.1109/TAFFC.2022.3172360.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8791–8800, June 2026.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25, 2018.

Chengxuan Qian, Shuo Xing, Li Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 152802–152824, 2026.

Arijit Ray, Ahmed Abdelkader, Chengzhi Mao, Bryan A Plummer, Kate Saenko, Ranjay Krishna, Leonidas Guibas, and Wen-Sheng Chu. Mull-tokens: Modality-agnostic latent thinking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9477–9488, 2026.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11329–11344, 2023.

Hao Sun, Ziwei Niu, Hongyi Wang, Xinyao Yu, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. Multimodal sentiment analysis with mutual information-based disentangled representation learning. *IEEE TAFFC*, 16(3):1606–1617, 2025. doi: 10.1109/TAFFC.2025.3529732.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE TAFC*, 15(1):309–325, 2023.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 5100–5111, 2019.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *ACL*, pp. 6558–6569, Florence, Italy, July 2019a. ACL. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019b.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE TMM*, 25:4909–4921, 2022.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis. *PR*, 136:109259, 2023.

Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. *Advances in Neural Information Processing Systems*, 37:66383–66409, 2024.

Yan Wang, Fuyuan Cao, and Xingwang Zhao. Conflict-aware adaptive cross-reconstruction for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15880–15889, June 2026.

Yiyu Wang, Haifang Jian, Jian Zhuang, Huimin Guo, and Yan Leng. Sslmm: Semi-supervised learning with missing modalities for multimodal sentiment analysis. *Information Fusion*, 120: 103058, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837, 2022.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *AAAI*, volume 39, pp. 1601–1609, 2025.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms. In *International Conference on Learning Representations*, volume 2026, pp. 127108–127138, 2026.

Heng Xie, Kang Zhu, Zhengqi Wen, Jianhua Tao, Xuefei Liu, Ruibo Fu, and Changsheng Li. Psamf: Personality-sentiment aligned multi-level fusion. In *AAAI*, volume 40, pp. 2191–2199, 2026.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *ACL*, pp. 7617–7630, Toronto, Canada, July 2023a. ACL. doi: 10.18653/v1/2023.acl-long.421. URL https://aclanthology.org/2023.acl-long.421/.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *ACL*, pp. 7617–7630, 2023b.

Zhilu Yang and Mingcheng Li. Factorize, reconstruct, enhance: A unified framework for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15860–15869, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Wengao Ye, Yan Liang, and Lianlei Shan. Thinking on the fly: Test-time reasoning enhancement via latent thought policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 156213–156239, 2026.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *ACL*, pp. 3718–3727, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*, volume 35, pp. 10790–10797, 2021.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *EMNLP*, pp. 1103–1114, Copenhagen, Denmark, Sept. 2017. ACL. doi: 10.18653/v1/D17-1115.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, volume 32, 2018a.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, pp. 2236–2246, 2018b.

Haoyu Zhang, Yinan Zhang, Chaolong Ying, Xiaoying Tang, and Tianshu Yu. Improving task-specific multimodal sentiment analysis with general mllms via prompting. In *NeurIPS*, 2026.

Xiangmin Zhang, Wei Wei, and Shihao Zou. Modal feature optimization network with prompt. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4611–4621, 2025.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356, 2024.

Cangzhi Zheng, Junjie Peng, Lan Wang, Li'an Zhu, Jiatao Guo, and Zesu Cai. Frame-level nonverbal feature enhancement based sentiment analysis. *Expert Systems with Applications*, 258:125148, 2024.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Denny Zhou, Swaroop Mishra, Huaixiu S Zheng, et al. Self-discover: Large language models self-compose reasoning structures. *Advances in Neural Information Processing Systems*, 37:126032–126058, 2024.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Fuji Ren. Kebr: Knowledge enhanced self-supervised balanced representation. In *ACM MM*, pp. 5732–5741, 2024.

Aoqiang Zhu, Min Hu, and Yan Xing. Fast retrieval and slow reasoning for explainable multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30381–30391, San Diego, California, United States, July 2026. ACL. ISBN 979-8-89176-395-1.

Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. Glomo: Global-local modal fusion for multimodal sentiment analysis. In *ACM MM*, pp. 1800–1809, 2024.

A APPENDIX

This appendix provides additional theoretical and empirical support for RDD. Section A.1 motivates the PID-inspired cross-modal roles, Section A.2 explains their realization in IGED, Section A.3 analyzes nonlinear synergy modeling, and Section A.4 interprets UALD as stochastic latent optimization. Further empirical analyses are provided in the subsequent sections. Section A.5 examines auxiliary-loss sensitivity, Section A.6 evaluates internal components, Section A.7 presents qualitative cases, Section A.8 compares key components with conventional alternatives, Section A.9 studies expert-pool capacity, and Section A.10 summarizes the datasets and evaluation protocols.

A.1 PID-INSPIRED CROSS-MODAL ROLES

Partial Information Decomposition (PID) (Williams & Beer, 2010) provides an information-theoretic lens for decomposing the information that two source variables X_1, X_2 jointly carry about a target Y. For the two-source case, the joint mutual information admits a unique decomposition as follows:

$$I(X_1, X_2; Y) = \mathrm{Red}(X_1, X_2; Y) + \mathrm{Syn}(X_1, X_2; Y) + \mathrm{Uniq}(X_1; Y) + \mathrm{Uniq}(X_2; Y), \quad (7)$$

where $\mathrm{Red}()$ is the information redundantly carried by both sources, $\mathrm{Syn}()$ is information available only through their interaction, and $\mathrm{Uniq}(X_1; Y)$ and $\mathrm{Uniq}(X_2; Y)$ are the information exclusive to X_1 and X_2, respectively.

In MSA the sources are the textual, acoustic, and visual streams, i.e., a three-source problem $X_1 = T, X_2 = A, X_3 = V$. The two-source identity above does not extend verbatim: for three sources the PID lattice contains additional atoms (e.g., PID-specific information and lower-order synergies), and no universally agreed decomposition into a single Red/Syn/Uniq triple exists. We therefore do *not* claim that our three expert channels recover an exact PID decomposition of $I(T, A, V; Y)$.

Instead, we adopt Redundancy, Synergy, and Uniqueness as "functional inductive biases" that correspond to three qualitatively distinct ways in which affective evidence can arise across modalities, which we term **cross-modal roles**.

Instead, we adopt Redundancy, Synergy, and Uniqueness as *functional inductive biases* that correspond to three qualitatively distinct ways in which affective evidence can arise across modalities:

- **Redundancy** – affective cues consistently conveyed by multiple modalities, which can serve as a stable consensus signal (e.g., an approving sentence accompanied by a smile and a pleasant tone).
- **Synergy** – affective meaning that emerges only from the joint configuration of modalities and is invisible to any single stream (e.g., sarcasm, where literal text, prosody, and facial expression jointly signal an opposite sentiment).
- **Uniqueness** – affective information carried by a single modality and absent from the others (e.g., a subtle prosodic cue not reflected in the lexical content).

MIRR instantiates these roles as a per-instance routing distribution $\mathbf{p} = (\pi_R, \pi_S, \pi_U)$, which weights the three specialized experts in IGED according to the cue configuration of each utterance.

A.2 CROSS-MODAL ROLE REALIZATION IN IGED

IGED instantiates the three cross-modal roles as dedicated experts, each shaped by an architecture and training objective that directly targets the corresponding type of affective evidence.

First, the $\mathcal{R}$-expert compresses each modality into a shared Gaussian latent, regularized by a KL term and pulled across modalities by a contrastive alignment loss. The dual design has a clear reading: the KL bottleneck discards variance that is not common to all three streams, while the contrastive loss enforces that the same utterance's textual, acoustic, and visual embeddings lie close to one another. What survives is the affective consensus–an approving wording, a pleasant tone, and a smile that mutually confirm each other–rather than the modality-specific fluctuations riding on top of it.

Second, the $\mathcal{S}$-expert lets each modality attend to the other two before a nonlinear projection combines them. Because the intermediate representation is formed only after cross-modal exchange, the expert cannot fall back on unimodal cues: any pattern it retains must arise from the joint configuration of the three streams. This is precisely the regime where sarcasm, irony, and conflicting cues live–cases in which literal text carries one sentiment while prosody and facial expression signal another. A KAN with learnable one-dimensional bases is placed at the projection stage because these interaction patterns are highly nonlinear and non-smooth, and a fixed MLP inductive bias tends to over-smooth them.

Then, the $\mathcal{U}$-expert explicitly conditions on what the $\mathcal{R}$-expert already explains. A dynamic gate uses F_{red}^m to suppress the part of x_m overlapping with the shared consensus, leaving only the residual; a 1D temporal convolutional network then summarizes this residual, and a reconstruction loss asks the unique and redundancy branches together to recover the original input. The gate prevents the unique branch from re-learning the consensus, while the reconstruction loss prevents it from collapsing to zero. The branch therefore specializes on modality-exclusive signals–for example, a subtle prosodic contour that the text alone does not convey.

Finally, the three experts are then combined by the MLLM-derived routing score $\beta = (\beta_R, \beta_S, \beta_U)$ from MIRR, so that each utterance is routed toward the information role matching its own cue configuration. The auxiliary losses $\mathcal{L}_{KL}$, $\mathcal{L}_{align}$, and $\mathcal{L}_{recon}$ jointly encourage the expert channels to specialize on disjoint affective factors, and Fig. 6 visualizes the resulting separation on held-out data.

A.3 Nonlinear Interaction Modeling for Synergy

The $\mathcal{S}$-expert's nonlinear projection is implemented as a Kolmogorov–Arnold Network (KAN), whose parameterization differs from a conventional MLP in one important way: instead of placing fixed activation functions on the nodes and learnable weights on the edges, a KAN places learnable one-dimensional activation functions on every edge and sums them at each node,

$$y_i^{(l+1)} = \sum_j \phi_{j,i}^{(l)}\left(W_{j,i}^{(l)}\, y_j^{(l)}\right), \tag{8}$$

where each $\phi_{j,i}^{(l)}$ is a spline-based function learned end to end. This edge-wise parameterization gives the network direct control over the shape of each univariate nonlinearity, rather than bending a fixed ReLU/GELU curve through affine projections.

The interaction patterns that matter for synergy–sarcasm, incongruity, mixed signals–are typically non-smooth and highly sample-dependent: a word that reads neutrally in isolation becomes hostile when paired with a flat affect, and the boundary between "polite disagreement" and "sarcastic agreement" is not captured by a globally smooth nonlinearity. Fixed MLP activations tend to average over such sharp transitions, whereas a KAN can dedicate one edge function to each relevant pairwise or higher-order combination and fit its shape on the data.

By construction, the synergy branch is invoked only after the cross-modal pooling step, so its input h_{pool} already encodes the joint configuration of the three streams. Combined with the routing score β_S, this means F_{syn} is asked to model what the redundancy branch has not already explained–an interaction residual rather than a fresh re-encoding of all three streams. The redundancy channel first absorbs the mutually consistent consensus; the synergy channel then fills in the higher-order, configuration-specific terms that remain. This residual reading keeps the two branches from competing for the same signal, and lets β_S scale the synergy contribution up exactly when the input contains joint cues that no single modality conveys.

Concretely, each KAN layer consists of two parallel branches: a residual MLP branch $\mathrm{base}(\cdot) = \mathrm{Linear}(\mathrm{SiLU}(\cdot))$ and a spline branch parameterized by fixed sinusoidal basis functions $b_g(x) = \sin(gx)$ for $g = 1, \ldots, G$ ($G = 5$). The spline output is

$$\mathrm{spline}(x)_o = \sum_{i=1}^{d}\sum_{g=1}^{G} w_{o,i,g}\, b_g(x_i), \tag{9}$$

where $W \in \mathbb{R}^{d\times d\times G}$ is a learnable tensor and $d = 256$. The layer output is $\mathrm{base}(x) + \mathrm{spline}(x)$. Two such layers are stacked in the $\mathcal{S}$-expert, both operating on $h_{pool} \in \mathbb{R}^{256}$.

A.4 UALD AS STOCHASTIC LATENT OPTIMIZATION

The refinement step in UALD can be read as a zero-order optimization over the latent representation z, rather than as a generic augmentation heuristic. Starting from the projected initial state $z^0 = \mathrm{Proj}(F_{\mathrm{fused}})$, the model samples B perturbations $\epsilon^b \sim \mathcal{N}(0, \sigma^2 I)$, evaluates the predictive confidence α_t^b at each perturbed point $z_b^t = z^t + \epsilon^b$, and takes the step

$$z^{t+1} = z^t + \lambda \sum_{b=1}^{B} w_t^b \, (z_b^t - z^t), \qquad w_t^b = \frac{\alpha_t^b}{\sum_{j=1}^{B} \alpha_j^t}. \tag{10}$$

This is a confidence-weighted finite-difference update: perturbations that land on points the predictor finds more plausible receive larger weight, so the expected displacement $\mathbb{E}_\epsilon[w(z^t+\epsilon) - w(z^t)]$ points toward regions of latent space the model already treats as high-confidence. No gradient of the loss with respect to z is required; the predictor's own confidence supplies the search direction.

The refinement loop is gated by predictive entropy. At each step t, the B candidate predictions are treated as samples from a Gaussian $\mathcal{N}(\mu_t, \sigma_t^2)$, where $\mu_t = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_b^t$ and $\sigma_t^2 = \frac{1}{B}\sum_{b=1}^{B}(\hat{y}_b^t - \mu_t)^2$. The differential entropy $\mathcal{H}(z^t) = \frac{1}{2}\log(2\pi e\, \sigma_t^2)$ is then normalized to $[0, 1]$ against the dataset-level label variance. When the entropy falls below the threshold γ, the refinement converges and the current state f^{LTT} is used for final prediction. Iteration is spent only on ambiguous or conflicting samples–exactly where a single forward pass is likely to err. Easy samples converge quickly and are predicted in few steps, while hard samples receive several rounds of refinement.

Confidence-weighted iteration has a blind spot: if the predictor is confidently wrong, UALD only pushes z deeper into the wrong high-confidence region. To catch these cases, the feedback queue flags samples that have stopped deliberating (low entropy) yet still show a large prediction error with high MLLM confidence π_{LLM}. This mechanism is applied only during training, as it requires ground-truth labels to identify erroneous predictions. These flagged samples are re-evaluated offline by the MLLM, which revises their cross-modal role vectors so that similar routing mistakes are less likely to recur.

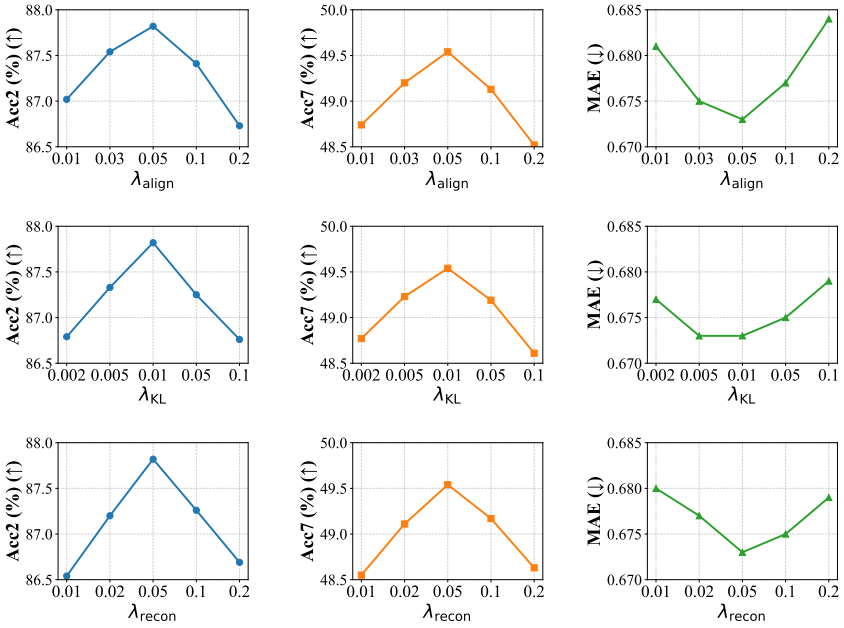


Figure 7: Sensitivity analysis of the three auxiliary weights on CMU-MOSI.

A.5 Auxiliary Loss Weight Sensitivity

The total training objective combines the task loss with three auxiliary regularizers $\mathcal{L}_{align}$, $\mathcal{L}_{KL}$, and $\mathcal{L}_{recon}$, weighted by λ_{align}, λ_{KL}, and λ_{recon}, respectively. To verify that the reported results are not sensitive to these weights, we sweep each weight individually on CMU-MOSI while fixing the other two at their default values. All other hyperparameters, the training schedule, and the random seed are kept identical to the main model. We report the standard test metrics (Acc_2, Acc_7, MAE) together with the fraction of runs that stably converge without collapsing.

As shown in Fig. 7, we sweep each auxiliary weight individually on CMU-MOSI while fixing the other two at their default values. Across all three sweeps, every metric peaks at the default configuration ($\lambda_{\text{align}} = 0.05$, $\lambda_{\text{KL}} = 0.01$, $\lambda_{\text{recon}} = 0.05$), where Acc_2 reaches 87.82, Acc_7 reaches 49.54, and MAE drops to 0.673. Varying each weight within roughly one order of magnitude on either side of the default changes Acc_2 by at most ± 0.8 points, Acc_7 by at most ± 0.5 points, and MAE by no more than 0.005. Even at the extremes of the swept range the performance degrades only marginally, and no training run exhibits collapse, indicating that the reported results are robust rather than relying on an overly narrow hyperparameter sweet spot. The three auxiliary terms therefore behave as well-behaved regularizers: moderate values stably align the latent space, regularize the posterior, and preserve reconstruction information, whereas excessive weights only slightly distort the task-specific objective without destabilizing optimization.

A.6 Internal Component Ablations

We further ablate internal components within each module. For the IGED expert pool, the three experts correspond to Redundancy ($\mathcal{R}$), Synergy ($\mathcal{S}$), and Uniqueness ($\mathcal{U}$). Whenever a specific expert is removed, the corresponding state vector in the preceding MLLM chain-of-thought stage is also excluded, ensuring that the routing state space strictly aligns with the available physical operators. For the feedback mechanism, w/o FB retains the UALD refinement but disables the offline error-sample queue and the MLLM re-evaluation loop during training.

Table 4: Internal component ablations on MOSI.

Setting	Acc2↑	F1↑	Acc7↑	MAE↓	Corr↑
w/o $\mathcal{R}$-expert	86.55	86.54	48.12	0.699	0.770
w/o $\mathcal{S}$-expert	87.20	87.17	48.55	0.689	0.779
w/o $\mathcal{U}$-expert	86.63	86.65	48.07	0.695	0.772
w/o FB	86.97	86.95	48.64	0.685	0.776
Full model	**87.82**	**87.79**	**49.54**	**0.673**	**0.825**

As shown in Table 4, all components are essential. Removing the $\mathcal{R}$-expert causes the largest drop (Acc_2 down 1.27%, MAE rising to 0.699), indicating that cross-modal invariant consensus serves as the primary anchor for stable prediction. Excluding the $\mathcal{U}$-expert yields the lowest Acc_7 (48.07%), showing that purified modality-specific nuances carry indispensable discriminative signals. Removing the $\mathcal{S}$-expert only mildly harms Acc_2 (87.20%) but notably degrades Acc_7 (down 0.99%) and Corr (0.779 vs. 0.825), demonstrating that non-linear cross-modal collisions are vital for resolving conflicting cues such as sarcasm. Finally, disabling the feedback loop (w/o FB) degrades all metrics (Acc_2 down 0.85%, Corr down 0.049), confirming that re-evaluating overconfident routing failures mitigates error accumulation over training.

A.7 Additional Case Study Analysis

In addition to the cross-modally conflicting sample analyzed in the main text, we report an additional qualitative case, as shown in Fig. 8.

Sample: Text-dominant negative case. The utterance (*"Boring..."*) is a short, strongly negative textual utterance. Visually, the speaker remains relatively neutral with only a faint frown (*Pred(V)*=$(+0.11)$), and auditorily the tone is flat rather than clearly negative (*Pred(A)*=$(+0.50)$); in other words, the non-verbal channels act as near-zero noise and carry little sentiment-relevant signal on their own, while the text alone conveys the dominant negative judgment. The MIRR prior therefore assigns the highest weight to Uniqueness ($(U = 0.42)$, versus $R = 0.30$ and $S = 0.28$), indicating that the text modality contributes a modality-unique signal that the visual and audio channels cannot provide. Following this routing, the fused prediction (-2.21) aligns with the text-dominant negative direction and closely matches the ground-truth label (-2.40), showing that MIRR correctly identifies which modality carries the decisive evidence when the others are uninformative.

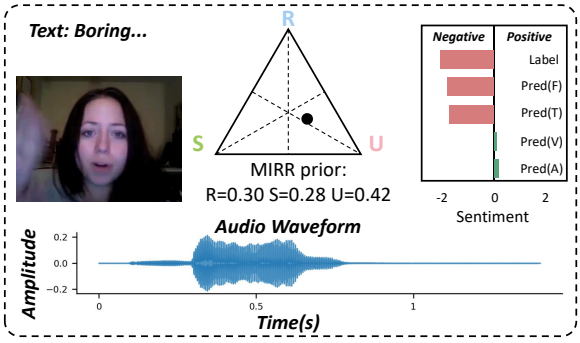


Figure 8: Additional qualitative case: text-dominant negative sample.

A.8 Key Component Replacements

To verify that RDD's design choices are supported by explicit architectural motivations rather than engineering preference, we replace three key components with classical alternatives while keeping all other settings identical. In **w/ MLP (no KAN)**, the KAN spline projection in the $\mathcal{S}$-expert is replaced with a standard two-layer GELU MLP, testing whether spline basis functions are necessary for discontinuous cross-modal interactions such as sarcasm. In **w/ Cross-modal Attn.**, the β-weighted expert aggregation is replaced with the crossmodal attention, testing whether cross-modal role routing outperforms generic attention. In **w/ Uniform Fusion**, the MLLM-derived β weights are replaced with fixed $1/3$ across the three experts, isolating the contribution of instance-specific cross-modal role vectors from the expert architecture itself.

Table 5: Key component replacements on MOSI.

Setting	Acc2↑	F1↑	Acc7↑	MAE↓	Corr↑
w/ MLP (no KAN)	87.26	87.21	48.67	0.685	0.781
w/ Cross-modal Attn.	86.85	86.85	48.17	0.697	0.779
w/ Uniform Fusion	86.17	86.15	47.34	0.705	0.772
Ours	**87.82**	**87.79**	**49.54**	**0.673**	**0.825**

As shown in Table 5, all three replacements consistently degrade performance. Replacing KAN with a standard MLP (**w/ MLP**) causes the smallest drop (Acc_2 down 0.56%, Corr 0.781), indicating that the spline nonlinearity brings a certain degree of impact on cross-modal interactions. Substituting crossmodal attention for cross-modal role routing (**w/ Cross-modal Attn.**) causes a larger drop (Acc_2 down 0.97%, MAE rising to 0.697), showing that MLLM-derived β weights provide stronger inductive bias than generic attention. Removing the instance-specific priors entirely (**w/ Uniform Fusion**) is the most detrimental (Acc_2 down 1.65%, Acc_7 down 2.20%, Corr falling to 0.772), confirming that MIRR-derived cross-modal role vectors are the most critical component—the expert architecture alone cannot recover the same discriminative signal without sample-specific routing.

A.9 Expert Pool Size Sensitivity

We further investigate how the number of $\mathcal{R}$-, $\mathcal{S}$-, $\mathcal{U}$- expert (RSU) triplets per modality affects performance. We sweep the number of RSU groups from 1 to 5 for the visual pool and from 1 to 4 for the text and audio pools, while keeping all other hyperparameters fixed at their default values. Each RSU group consists of one redundancy, one synergy, and one uniqueness expert, so the total expert count per modality is $3N_m$.

As shown in Fig. 9, performance is relatively stable across a reasonable range of expert pool sizes. The visual pool benefits from a moderately larger capacity ($N_V = 3$, i.e., 9 visual experts), which is expected given that visual information is richer and requires more specialized operators. For the text and audio pools, $N_T = N_A = 2$ (i.e., 6 experts each) yields the best trade-off; further increasing the expert count leads to slight performance degradation, likely due to over-parameterization on the smaller MOSI training set. Overall, the model is robust to moderate variations in pool size, and the default configuration ($N_T = 2$, $N_V = 3$, $N_A = 2$) consistently achieves the best results.

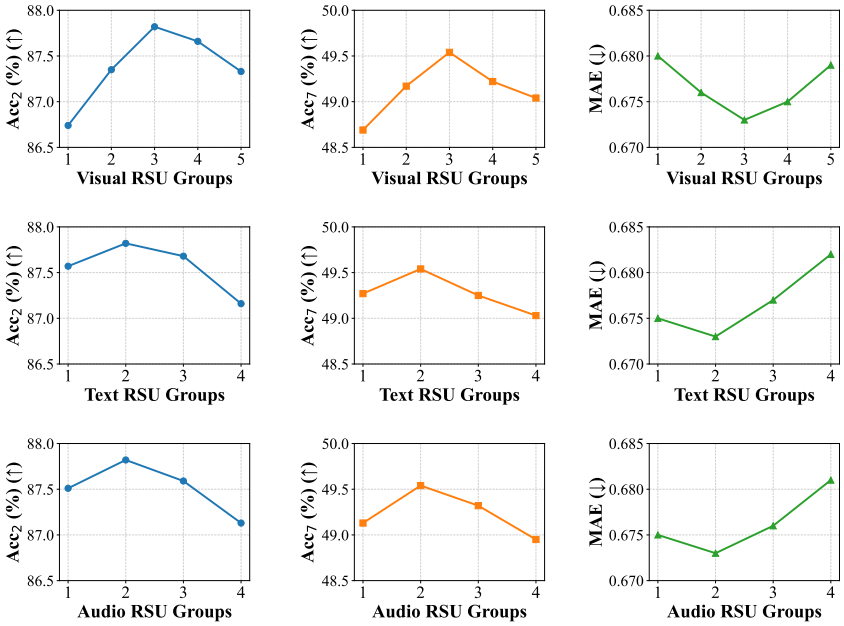


Figure 9: Sensitivity analysis of expert pool size per modality on CMU-MOSI.

A.10 Experiments Dataset

CMU-MOSI contains 2,199 English opinion clips (split into 1,284/229/686 for training/validation/testing), while CMU-MOSEI scales this up to 23,453 clips (16,326/1,871/4,659). Both utilize a continuous sentiment annotation system ranging from -3 to +3. CH-SIMS provides a Chinese language benchmark with 2,281 multimodal samples (1,368/456/457). We adopt standard evaluation metrics across these datasets, including MAE, Corr, Acc-7, Acc-2, and F1.