RIDER-MoE: Reliability-Aware Expert Routing for Robust Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Multimodal sentiment models are typically developed and evaluated on curated benchmarks where text, audio, and vision are well-aligned and reliable. In deployment, modality quality varies across instances (e.g., noise, missing sensors), and modalities can disagree, making fixed fusion strategies brittle. We present **RIDER-MoE**, a mixture-of-experts architecture that routes each example among modality-uniqueness, redundancy, and synergy experts based on estimated modality reliability and cross-modal agreement. The expert decomposition is motivated by Partial Information Decomposition (PID), and we operationalize the intended U/R/S semantics via masked-view *Disentangled Interaction Regularization* during training. The router augments the fused representation with lightweight unimodal sentiment probes: high probe entropy and low consensus down-weight synergistic fusion and shift probability mass toward redundancy or unimodal experts. On CMU-MOSI and CMU-MOSEI, RIDER-MoE is competitive with recent strong baselines on clean test sets and achieves the best robustness (highest normalized AUC) across noise, missing-modality, and cross-modal conflict stress tests. These results support reliability-aware expert routing as a practical mechanism for robust multimodal sentiment analysis. Code: `https://anonymous.4open.science/r/submission-5B6D`

1 Introduction

Human communication is a multifaceted process where sentiment is conveyed through the intricate interplay of linguistic content, acoustic resonance, and facial expressions (Poria et al., 2023; Li et al., 2024). The field of Multimodal Sentiment Analysis (MSA) has witnessed a surge in performance, with recent deep learning architectures (Zhou et al., 2025; Fang et al., 2025; He et al., 2025; Li and Li, 2025) achieving impressive accuracy on benchmarks such as MOSI and MOSEI. These SOTA models excel at mining subtle cross-modal dependencies, leveraging the assumption that complex fusion yields superior understanding.

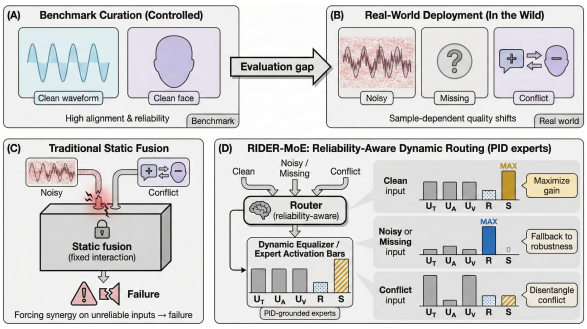

Figure 1: Benchmark vs. deployment: static fusion can fail under sample-dependent modality unreliability; reliability-aware routing adapts expert usage.

However, this "benchmark prosperity" often masks a critical vulnerability. As illustrated in **Figure 1 (Top)**, standard benchmarks function (Zadeh et al., 2016; Bagher Zadeh et al., 2018) as "greenhouses"—curated environments characterized by high alignment, clear signals, and minimal noise. In stark contrast, real-world deployment is fraught with sample-dependent quality shifts: audio may be corrupted by background chatter, faces may be occluded, and modalities may exhibit semantic conflicts (e.g., sarcasm where positive text contradicts a negative tone) (Poria et al., 2020).

Most existing approaches implicitly adopt a *static interaction* assumption: the model learns one dominant cross-modal interaction pattern and applies it to all samples (Wu et al., 2025; Chen et al., 2025). As depicted in the "Static Collapse" scenario of **Figure 1 (Bottom-Left)**, this amounts to a *globally fixed fusion strategy shared across samples* that aggressively seeks *Synergy* regardless of per-sample reliability. When a modality becomes unreliable (e.g., noise, occlusion, missing sensors), forcing synergistic fusion entangles corrupted evidence with clean cues and can con-

taminate predictions; when modalities conflict, the same aggressive fusion can wash out informative modality-specific signals. Robust MSA therefore requires *sample-wise* strategy selection, together with training signals that keep different interaction pathways behaviorally distinct rather than collapsing into a single undifferentiated route.

To address this, we propose **RIDER-MoE** (**R**eliability-aware **I**nteraction-**D**isentangled **E**xpert **R**outing), a tri-modal mixture-of-experts model that adapts its fusion strategy on a per-sample basis (**Figure 1, Bottom-Right**). RIDER-MoE routes each example among five experts: three unimodal *uniqueness* experts (U_T, U_A, U_V), a *redundancy* expert (R) designed to remain stable under single-modality degradation, and a high-gain but fragile *synergy* expert (S). This expert factorization is *inspired* by Partial Information Decomposition (PID) (Williams and Beer, 2010; Wollstadt et al., 2023); importantly, we do not estimate PID quantities or commit to a specific redundancy definition, but instead operationalize the intended U/R/S semantics via masked-view Disentangled Interaction Regularization (DIR; Appendix A). Finally, a **Reliability-Aware Router** that monitors each modality's uncertainty (via entropy) and cross-modal agreement (via a consensus score) and uses these signals to adjust the expert mixture *per sample*. For **clean and consistent** inputs, the router upweights the *Synergy* expert to exploit interaction-only cues for maximal information gain and precision. For **noisy or missing** modalities, elevated uncertainty (or absence) suppresses the fragile Synergy pathway and shifts probability mass toward the *Redundancy* expert, promoting stable predictions under modality degradation. For **cross-modal conflicts**, agreement drops even when unimodal confidence remains high (e.g., positive text paired with negative audio), so the router emphasizes the corresponding *Uniqueness* experts to preserve reliable modality-specific evidence and avoid cancellation from contradictory signals.

Our contributions are:

- **Reliability-aware expert routing for MSA.** We introduce RIDER-MoE, which dynamically routes examples between synergy-, redundancy-, and modality-specific experts based on uncertainty and cross-modal consensus signals.
- **Masked-view interaction disentanglement.** We propose Disentangled Interaction Regularization (DIR), a masked-view regularizer that enforces uniqueness/redundancy invariances and constrains synergy to act as a full-view residual, promoting stable expert roles.
- **Robustness evaluation and results.** On MOSI and MOSEI, RIDER-MoE is competitive on clean benchmarks and yields consistent robustness gains under noise, missing-modality, and conflict stress tests, supported by ablations.

2 Related Work

Multimodal sentiment analysis. MSA has progressed from early tensor/low-rank fusion (Morency et al., 2011; Zadeh et al., 2017) to attention-based cross-modal interaction modeling (Poria et al., 2023; Tsai et al., 2019a). Recent strong architectures include dual-path fusion (Zhou et al., 2025), state-space models for long sequences (He et al., 2025), and parameter-efficient adapters in PLMs (Chen et al., 2025). Many methods still rely on static fusion/gating at test time (Wu et al., 2025), which can fail under modality-quality shifts and conflicts (Poria et al., 2020). RIDER-MoE instead performs sample-wise expert routing based on estimated reliability and agreement.

Robustness via denoising and reconstruction. Robust MSA has been approached via denoising noisy modalities (Li and Li, 2025) or reconstructing missing ones (Zhu et al., 2025). These methods can incur extra computation and may assume a consistently reliable anchor modality. Our approach is complementary: rather than reconstructing inputs, RIDER-MoE adapts the interaction strategy by routing away from fragile synergy when probes indicate low reliability or low agreement.

Interaction disentanglement and MoE. PID characterizes multivariate information into unique, redundant, and synergistic terms (Williams and Beer, 2010; Wollstadt et al., 2023) and has been used to analyze multimodal model behavior (Liang et al., 2023). Separately, MoE offers conditional computation (Chen et al., 1999, 2023; Zhao et al., 2024) and has been explored in multimodal settings (e.g., (Fang et al., 2025; Xin et al., 2025)). Unlike conventional modality reweighting approaches, RIDER-MoE routes among *interaction-type* experts (U/R/S) and uses probe-based reliability plus

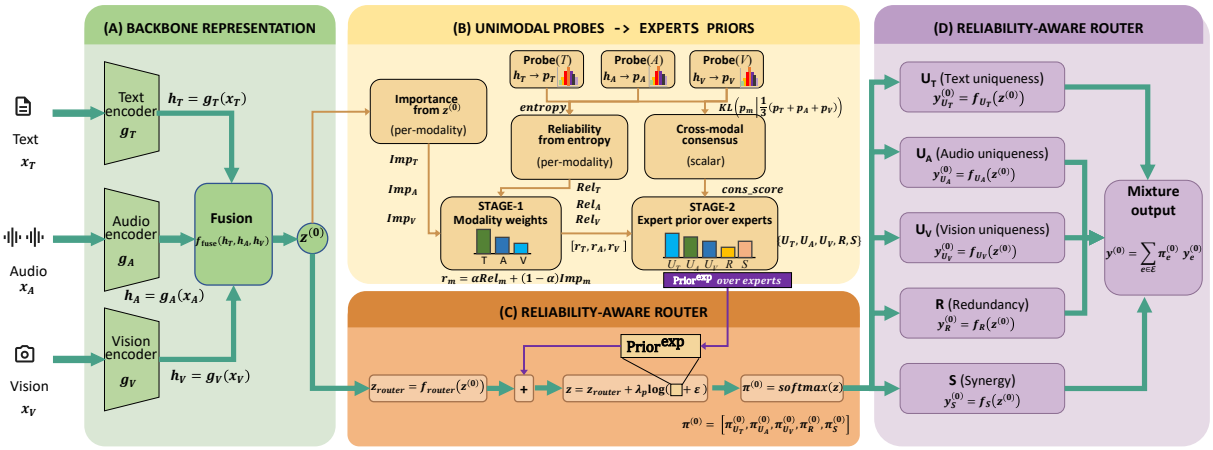

Figure 2: Inference-time architecture of RIDER-MoE with a reliability-aware router over *PID-inspired* U/R/S experts.

cross-modal consensus to bias routing; DIR further prevents expert-role collapse via masked-view constraints (Appendix A).

3 Method

We propose a tri-modal mixture-of-experts (MoE) architecture for sentiment analysis with text, audio, and vision (Figure 2). The model combines a reliability-aware router with interaction-specialized experts and a disentangled interaction regularization (DIR) scheme that shapes experts into uniqueness, redundancy, and synergy roles.

3.1 Problem Setup and Architecture Overview

We consider regression from tri-modal input (x_T, x_A, x_V) to a scalar sentiment label $y \in \mathbb{R}$ (e.g., in $[-3, 3]$). Let the modality set be $\mathcal{M} = \{T, A, V\}$. For each modality $m \in \mathcal{M}$, a unimodal encoder g_m maps the raw input x_m into a latent vector $h_m = g_m(x_m) \in \mathbb{R}^{d_m}$. A fusion module f_{fuse} combines the three encodings into a shared representation $z^{(0)} = f_{\text{fuse}}(h_T, h_A, h_V) \in \mathbb{R}^{d_z}$, which serves as the input to all experts and to the router on the full (unmasked) view, denoted by $p = 0$.

Interaction-aware experts. We instantiate a set of five scalar experts $\mathcal{E} = \{U_T, U_A, U_V, R, S\}$, where U_T, U_A, U_V are text-, audio-, and vision-uniqueness experts, R is a redundancy expert, and S is a synergy expert. For any view index p (full or masked; Section 3.3), each expert produces a scalar prediction $y_e^{(p)} = f_e\big(z^{(p)}\big) \in \mathbb{R}$, $e \in \mathcal{E}$.

Routing and prediction. On the full tri-modal view $p = 0$ (all modalities present), a reliability-aware router (Section 3.2) produces a probability vector over experts $\pi^{(0)} = \left[\pi_{U_T}^{(0)}, \pi_{U_A}^{(0)}, \pi_{U_V}^{(0)}, \pi_R^{(0)}, \pi_S^{(0)}\right]^\top \in \mathbb{R}^5$, with $\pi_e^{(0)} \geq 0$ and $\sum_{e\in\mathcal{E}} \pi_e^{(0)} = 1$. The routed prediction on the full view is

$$\hat{y}^{(0)} = \sum_{e\in\mathcal{E}} \pi_e^{(0)}\, y_e^{(0)}. \tag{1}$$

We use an absolute-error regression loss

$$\mathcal{L}_{\text{task}} = \left|\hat{y}^{(0)} - y\right|. \tag{2}$$

3.2 Reliability-Aware Routing

The router should not only depend on the fused representation $z^{(0)}$ but also account for how reliable and mutually consistent each modality appears on the current example. We therefore augment a standard MoE router with unimodal sentiment probes, entropy-based modality reliability, modality importance predicted from the fused context, and a cross-modal consensus score. These signals are aggregated into expert-level priors that softly bias the router logits. Full mathematical definitions are provided in Appendix B.1.

Unimodal sentiment probes. For each modality $m \in \mathcal{M}$, we attach a small classification head to h_m that predicts a discrete sentiment label $\tilde{y}$ obtained by quantizing y (e.g., to $K = 7$ levels). A softmax over the probe logits yields a unimodal sentiment distribution p_m and a cross-entropy loss $\mathcal{L}_{\text{cls}}^{(m)}$. These probes regularize the encoders and provide uncertainty-aware signals; they do not directly appear in Eq. (1).

Reliability, importance, and consensus. From each p_m we compute an entropy H_m and convert entropies into a relative reliability distribution Rel_m, where lower entropy yields higher reliability. In parallel, a linear layer on $z^{(0)}$ followed by softmax predicts modality importance scores Imp_m, capturing how useful each modality appears in the fused context. We blend these into normalized modality reliability weights r_m (and their average $\bar{r}$), which are later used to weight regularization terms. To measure cross-modal agreement, we compare each p_m to the mean distribution $\bar{p}$ using KL divergence and map the average disagreement into a scalar consensus score cons_score $\in (0, 1]$.

Expert priors and biased routing. The modality weights (r_T, r_A, r_V) and consensus score cons_score form a feature vector that is mapped, via a linear layer and softmax, to an expert prior distribution $\mathrm{Prior}^{\mathrm{exp}}$ over $\mathcal{E}$. In parallel, a router network f_{router} applied to $z^{(0)}$ produces base expert logits. We inject the expert prior as a log-bias before the final softmax, i.e., adding $\lambda_p \log(\mathrm{Prior}_e^{\mathrm{exp}})$ to the base logit of expert e, with strength $\lambda_p \geq 0$. This yields reliability- and consensus-aware routing probabilities $\pi^{(0)}$, while keeping the router fully trainable through $\mathcal{L}_{\text{task}}$.

3.3 Interaction Experts and Masked Views

DIR (Section 3.4) requires observing how expert outputs change when individual modalities are removed. To keep computation efficient, we reuse the unimodal encoders and manipulate their outputs via masking rather than re-encoding from scratch. After computing (h_T, h_A, h_V), we construct four views indexed by p: the full view $p = 0$, and three masked views $p \in \{-T, -A, -V\}$ where the corresponding modality representation h_m is replaced by a masking vector MASK_m in the same space as h_m. For each view $p \in \{0, -T, -A, -V\}$ we reuse the same fusion module and experts to obtain a fused representation $z^{(p)} = f_{\text{fuse}}\big(h_T^{(p)}, h_A^{(p)}, h_V^{(p)}\big)$ and scalar expert outputs $y_e^{(p)} = f_e\big(z^{(p)}\big)$ for all $e \in \mathcal{E}$. All routing-related quantities are computed once from the full view $(h_T, h_A, h_V, z^{(0)})$. The resulting $\pi^{(0)}$ is reused wherever routing weights are needed in the losses. This design avoids multiple router passes and ensures that all regularization shapes experts under a single routing policy derived from the complete tri-modal context.

3.4 Disentangled Interaction Regularization

DIR encourages the experts to specialize into modality uniqueness, redundancy, and synergy roles while leaving the main forward prediction Eq. (1) unchanged. It operates only on the scalar expert outputs $\{y_e^{(p)}\}$ across masked views and uses the reliability weights r_m and $\bar{r}$ from the router.

3.4.1 PID-inspired invariance for uniqueness and redundancy

The first component, L_{PID}, promotes invariance of uniqueness and redundancy experts across appropriate views, inspired by partial information decomposition. For each modality $m \in \{T, A, V\}$ we define the set of views where m is present, for example $P_T = \{0, -A, -V\}$, $P_A = \{0, -T, -V\}$, and $P_V = \{0, -T, -A\}$. The uniqueness expert U_m is encouraged to output similar values across all views in P_m, using a mean-squared deviation between $y_{U_m}^{(0)}$ and $y_{U_m}^{(p)}$ for $p \in P_m \setminus \{0\}$, weighted by the reliability r_m. If U_m encodes information from other modalities, its output changes when those modalities are masked, increasing the penalty.

The redundancy expert R is intended to capture information that persists when any single modality is removed, so we penalize the variance of $y_R^{(p)}$ across all four views $p \in \{0, -T, -A, -V\}$, scaled by the average reliability $\bar{r}$. The combined PID-inspired invariance loss is

$$\mathcal{L}_{\text{PID}} = \mathcal{L}_{\text{uni}} + \mathcal{L}_{\text{red}},$$

where $\mathcal{L}_{\text{uni}}$ and $\mathcal{L}_{\text{red}}$ denote the uniqueness and redundancy invariance terms, respectively. Exact formulas are given in Appendix B.3.

3.4.2 Synergy regularization

The second component, $\mathcal{L}_{\text{syn}}$, enforces a synergy interpretation: the synergy expert should be silent whenever any modality is missing and, on the full view, should explain the residual error left by the other experts.

First, when a modality is masked ($p \in \{-T, -A, -V\}$), we penalize non-zero synergy outputs $y_S^{(p)}$, using a squared penalty scaled by $\bar{r}^3$ to emphasize examples where all three modalities are reliable. This encourages the synergy expert to encode information that truly requires all three modalities simultaneously.

Second, on the full view $p = 0$, we view synergy as a residual correction. Let $\hat{y}_{\setminus S}^{(0)}$ be the routed

prediction obtained from all experts except S (using the same $\pi^{(0)}$). We construct a residual target $\delta = y - \mathrm{sg}(\hat{y}^{(0)}_{\setminus S})$, where $\mathrm{sg}(\cdot)$ is a stop-gradient operator, and encourage the routed synergy contribution $\pi^{(0)}_S y^{(0)}_S$ to match δ, again weighted by $\bar{r}^3$. Because gradients are stopped through $\hat{y}^{(0)}_{\setminus S}$, this term updates only the synergy expert and its routing weight. We denote the sum of the "off-under-masking" and "residual-fit" terms by $\mathcal{L}_{\text{syn}}$; detailed expressions are given in Appendix B.3.

3.5 Overall Objective and Optimization

3.5.1 Global loss

The overall loss for a batch combines the regression task, unimodal probe losses, and the two DIR components:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{cls}} \sum_{m \in \mathcal{M}} \mathcal{L}^{(m)}_{\text{cls}} + \lambda_{\text{PID}} \mathcal{L}_{\text{PID}} + \lambda_{\text{syn}} \mathcal{L}_{\text{syn}}. \tag{3}$$

Here, λ_{cls} weights the auxiliary probe losses, λ_{PID} controls the strength of the uniqueness and redundancy invariance constraints, and λ_{syn} controls synergy regularization. The scalar λ_p (Section 3.2) controls how strongly expert priors bias routing but does not enter the loss explicitly. DIR introduces only two global hyperparameters specific to interaction disentanglement (λ_{PID} and λ_{syn}).

3.5.2 Training vs. inference

During training we use the full and masked views, the auxiliary probe losses, and both DIR components, always routing with $\pi^{(0)}$ computed from the full view. At inference time we use only the full view and the routed prediction in Eq. (1); no masked views or DIR terms are involved. Thus the prediction cost matches that of a single-pass 5-expert MoE layer with a reliability-aware router (App. D). Detailed pseudo-code for training and inference is given in Algorithm 1 in Appendix B.4.

4 Experiments

Datasets, Metrics, and Baselines We conduct a comprehensive evaluation on two widely used public benchmarks, MOSI (Zadeh et al., 2016) and MOSEI (Bagher Zadeh et al., 2018). Following prior work, we report Acc-7, Acc-2, F1, MAE, and Corr. Implementation details (experimental setup, datasets, and hyperparameters), as well as the robustness stress-test protocols and the computation of robustness/diagnostic summary metrics, are provided in the Appendix C. We compare against representative state-of-the-art methods, including TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MulT (Tsai et al., 2019b), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), DMD (Li et al., 2023), EMOE (Fang et al., 2025), t-HNE (Li and Li, 2025), and MMA (Chen et al., 2025).

4.1 Main Results on Clean Benchmarks

Table 1 reports results on MOSI and MOSEI under the standard (clean) setting. On MOSEI, RIDER-MoE attains the highest Acc-7 among compared methods (55.3) and is comparable to the strongest clean baseline on Acc-2/F1 (85.67/85.61 vs. 85.7/85.7). On MOSI, RIDER-MoE achieves 47.13 Acc-7 and 85.12 Acc-2, which is within the performance range of recent competitive approaches (e.g., EMOE and t-HNE).

For regression-oriented metrics, RIDER-MoE obtains MAE/Corr of 0.714/0.796 on MOSI and 0.531/0.772 on MOSEI, remaining close to the best reported correlations in Table 1. Overall, these results indicate that the PID-inspired expert decomposition and the associated training regularization maintain competitive performance in the standard benchmark regime, which provides a clean-data reference point for the robustness evaluation in §4.2.

4.2 Robustness to In-the-Wild Quality Shifts

To focus the robustness comparison, we select the top five state-of-the-art baselines from §4.1 (MMIM, DMD, EMOE, t-HNE, and MMA).

Figure 3 reports performance–severity curves for three test-time stressors—Noise (ρ, within-modality corruption), Missingness (m, modality dropout), and Cross-modal Conflict (c, mismatched modalities that break cross-modal agreement)—on both MOSI and MOSEI. The exact corruption operators are defined in Appendix C.2. We sweep $\rho \in \{0, 0.1, \ldots, 0.5\}$, $m \in \{0, 0.2, \ldots, 0.8\}$, and $c \in \{0, 0.25, \ldots, 1.0\}$. As expected, all methods degrade as severity increases; however, RIDER-MoE shows a consistently slower decline, with the largest gaps emerging at moderate-to-high severities, especially under Missingness and Conflict.

To summarize robustness over the full severity range, Table 2 reports *normalized AUC*, computed by trapezoidal integration over the performance–severity curve and dividing by the severity range (equivalently, the average performance over the severity grid; Appendix C.3). On MOSI, RIDER-MoE achieves the highest normalized AUC on both

Methods	MOSI					MOSEI				
	ACC7↑	ACC2↑	F1↑	MAE↓	Corr↑	ACC7↑	ACC2↑	F1↑	MAE↓	Corr↑
TFN	34.9	80.8	80.7	0.901	0.698	50.2	82.5	82.1	0.593	0.700
LMF	33.2	82.5	82.4	0.917	0.695	48.0	82.0	82.1	0.623	0.677
MulT	35.1	80.2	80.1	0.936	0.711	52.3	82.7	82.8	0.572	0.753
MISA	41.8	84.2	84.2	0.754	0.766	52.3	85.3	85.1	0.543	0.723
Self-MM	45.3	84.9	84.9	0.738	0.789	53.2	84.5	84.3	0.540	0.758
MMIM	45.8	84.6	84.5	0.717	0.786	50.1	83.6	83.5	0.580	0.756
DMD	46.2	83.2	83.2	0.721	0.773	52.4	84.8	84.7	0.546	0.764
EMOE	**47.49**	85.13	84.89	0.710	0.804	53.91	85.3	85.3	0.537	0.782
t-HNE	47.04	85.02	84.98	**0.680**	**0.810**	54.05	85.2	85.32	**0.520**	**0.789**
MMA	46.90	**86.4**	**86.40**	0.693	0.803	55.2	**85.7**	**85.7**	0.529	0.766
RIDER-MoE (**Ours**)	47.13	85.12	85.12	0.714	0.796	**55.3**	85.67	85.61	0.531	0.772

Table 1: Main results on MOSI and MOSEI under the standard (clean) setting.

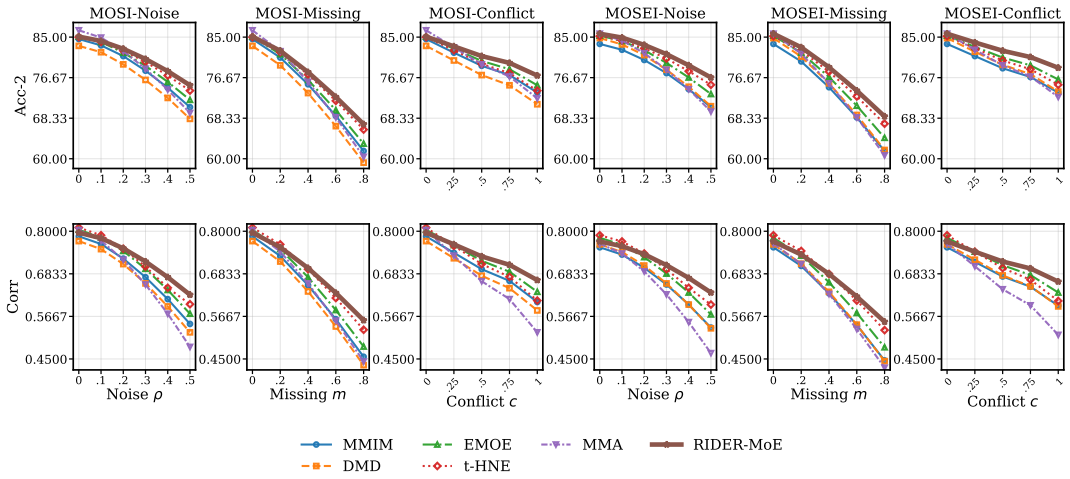

Figure 3: Robustness performance–severity curves under three test-time stressors: Noise (ρ), Missingness (m), and Cross-modal Conflict (c). Top: Acc-2. Bottom: Corr. Columns: MOSI/MOSEI under Noise/Missingness/Conflict. Stress-test protocols and severity grids are defined in Appendix C.2.

Acc-2 and Corr across all three stressors—Noise (81.1/0.727), Missingness (77.2/0.691), and Conflict (81.2/0.734). On MOSEI, RIDER-MoE shows the same consistent trend, attaining the top AUC under Noise (82.0/0.715), Missingness (78.2/0.676), and Conflict (82.3/0.719), indicating the strongest average robustness over the full severity range.

These results match RIDER-MoE's specialization and routing behavior. Under quality degradation (Noise/Missingness), the reliability-aware router suppresses the high-gain but brittle Synergy route and favors the Redundancy expert; under semantic disagreement (Conflict), it leans toward Uniqueness experts to mitigate cross-modal contamination. DIR further stabilizes redundancy/uniqueness under masked views and restricts synergy to a full-view residual correction, improving average Acc-2 and Corr across both quality shifts and disagreement.

4.3 Ablation Study

We conduct claim-aligned ablations on MOSI and MOSEI by removing *one* component (module or loss term) at a time while keeping the unimodal encoders, fusion backbone, optimization recipe, and inference procedure fixed. We ablate reliability-aware routing by (i) disabling expert-prior logit injection ($\lambda_p{=}0$) and (ii) removing the *consensus* term from the prior while retaining reliability-related terms; we ablate disentangled interaction regularization (DIR) by setting $\lambda_{\text{PID}}{=}\lambda_{\text{syn}}{=}0$ and by removing its sub-terms (L_{PID}, L_{syn}, $L_{\text{syn-off}}$, $L_{\text{syn-res}}$); we remove masked-view construction ($p{\in}0, -T, -A, -V$), eliminating view-based constraints; and we collapse the PID-motivated expert decomposition into a *single* prediction head, matching capacity by increasing the head width to approximate the parameter budget of the original multi-expert heads. Table 3 reports clean metrics

Model	MOSI						MOSEI					
	Noise AUC↑		Missing AUC↑		Conflict AUC↑		Noise AUC↑		Missing AUC↑		Conflict AUC↑	
	Acc-2	Corr	Acc-2	Corr	Acc-2	Corr	Acc-2	Corr	Acc-2	Corr	Acc-2	Corr
MMIM	78.9	0.689	74.5	0.641	79.3	0.699	78.3	0.667	73.9	0.620	78.8	0.679
DMD	77.1	0.672	72.7	0.623	77.4	0.681	79.1	0.671	74.7	0.623	79.5	0.682
EMOE	79.9	0.711	75.4	0.663	80.3	0.722	80.4	0.697	76.1	0.650	81.0	0.710
t-HNE	80.5	0.719	76.6	0.687	79.7	0.714	81.2	0.709	77.3	0.674	80.4	0.702
MMA	79.5	0.673	75.0	0.645	79.7	0.668	79.3	0.645	74.7	0.616	79.4	0.646
RIDER-MoE (Ours)	**81.1**	**0.727**	**77.2**	**0.691**	**81.2**	**0.734**	**82.0**	**0.715**	**78.2**	**0.676**	**82.3**	**0.719**

Table 2: Robustness summary using *normalized AUC* over performance–severity curves for Acc-2 and Corr (higher is better). Normalized AUC is computed by trapezoidal integration and dividing by the severity range (Appendix C.3); stress-test protocols follow Appendix C.2.

and robustness nAUC. Additional hyperparameter sensitivity analyses are provided in Appendix C.6.

Table 3 shows that RIDER-MoE (Full) is consistently best overall on robustness across both datasets, while remaining competitive on clean benchmarks. Router prior ablations primarily reduce robustness with limited effect on clean metrics: disabling prior injection (λ_p=0) noticeably lowers Noise/Missing AUC on both Corr and Acc-2 (e.g., on MOSI, Noise-AUC(A2) drops from 81.1 to 79.3 and Missing-AUC(C) from 0.691 to 0.671), whereas removing only the consensus term leaves Noise/Missing closer to the full model but degrades conflict handling more sharply (e.g., Conflict-AUC(C) decreases from 0.734 to 0.704 on MOSI and from 0.719 to 0.685 on MOSEI). This pattern is consistent with the consensus signal being the key indicator for semantic disagreement, while reliability-related terms are more directly tied to noise/missingness.

DIR and view perturbations are required for broad robustness. Removing DIR entirely (λ_{PID}=λ_{syn}=0) produces consistent degradation on both clean and robustness metrics, especially under Missing (e.g., Missing-AUC(A2) decreases from 77.2 to 72.9 on MOSI and from 78.2 to 73.4 on MOSEI). The component ablations align with the intended roles: removing L_{PID} disproportionately harms conflict robustness (e.g., Conflict-AUC(C) 0.734→0.709 on MOSI), while removing L_{syn} mainly reduces Noise/Missing AUC. Within L_{syn}, removing $L_{\text{syn-off}}$ most strongly impacts Missing-AUC, whereas removing $L_{\text{syn-res}}$ more visibly degrades clean regression (e.g., on MOSI, MAE/Corr increases from 0.714/0.796 to 0.731/0.786). Training without masked views yields robustness drops close to removing DIR, indicating that the view-based constraints are necessary for DIR to shape expert behaviors. Finally, collapsing to a single expert yields the largest overall robustness degradation (e.g., Missing-AUC(A2) drops to 71.2/MOSI and 70.8/MOSEI), supporting the necessity of the PID-motivated U/R/S decomposition together with reliability-aware routing for robust multimodal sentiment prediction.

4.4 Analysis: Routing Dynamics and Interaction Disentanglement

Routing dynamics under quality shifts and conflicts. Figure 4 plots the average full-view routing weights $\mathbb{E}[\pi^{(0)}]$ as corruption severity increases, where $\pi_U = \pi_{U_T} + \pi_{U_A} + \pi_{U_V}$. On clean inputs, the router assigns most mass to synergy (π_S). As Noise (ρ) or Missingness (m) increases, π_S decreases while redundancy (π_R) rises, with a stronger shift under Missingness; π_U stays relatively small. This matches the router's *entropy-based modality reliability* and *modality importance* signals and their *log-prior injection*: when a modality becomes uncertain/absent, the injected expert prior down-weights the fragile synergistic route and favors the redundancy expert, whose semantics are designed to persist under single-modality degradation. Under Conflict (c), the trend changes: π_U increases substantially as π_S drops, while π_R stays near its baseline. Mechanistically, conflict primarily reduces the *cross-modal consensus score* (without necessarily reducing unimodal confidence), so the consensus-conditioned prior shifts mass toward uniqueness experts to avoid cross-modal contamination in a synergistic fusion.

DIR semantic checks and ablation signatures. Figure 5 evaluates whether DIR enforces interpretable U/R/S roles using masked views. We report four behavioral diagnostics aligned with DIR: (i) *Uniqueness consistency* ΔU across views where

Variant	MOSI											MOSEI										
	Clean					Robust AUC						Clean					Robust AUC					
	A7	A2	F1	MAE	C	N(C)	M(C)	Cf(C)	N(A2)	M(A2)	Cf(A2)	A7	A2	F1	MAE	C	N(C)	M(C)	Cf(C)	N(A2)	M(A2)	Cf(A2)
RIDER-MoE (Full)	**47.13**	**85.12**	**85.12**	**0.714**	**0.796**	**0.727**	**0.691**	**0.734**	**81.1**	**77.2**	**81.2**	**55.3**	**85.67**	**85.61**	**0.531**	**0.772**	**0.715**	**0.676**	**0.719**	**82.0**	**78.2**	**82.3**
w/o prior injection (λ_p=0)	46.93	84.82	84.80	0.720	0.792	0.709	0.671	0.724	79.3	75.0	80.2	55.1	85.45	85.38	0.535	0.768	0.696	0.654	0.707	80.0	75.8	81.1
w/o consensus term in prior	47.03	84.92	84.90	0.718	0.793	0.721	0.683	0.704	80.5	76.4	78.3	55.2	85.50	85.44	0.534	0.769	0.709	0.668	0.685	81.2	77.2	79.0
w/o DIR (λ_{PID}=λ_{syn}=0)	46.33	84.22	84.18	0.734	0.784	0.694	0.652	0.700	77.8	72.9	78.6	54.4	84.70	84.62	0.548	0.758	0.684	0.635	0.688	78.4	73.4	79.4
w/o L_{PID}	46.63	84.52	84.50	0.726	0.788	0.715	0.678	0.709	79.9	75.6	78.9	54.8	85.05	84.98	0.540	0.763	0.704	0.662	0.693	80.9	76.6	79.6
w/o L_{syn} (remove $L_{\text{syn-off}}$ & $L_{\text{syn-res}}$)	46.73	84.62	84.60	0.724	0.790	0.707	0.664	0.717	79.1	74.6	79.8	54.9	85.10	85.02	0.539	0.765	0.694	0.646	0.706	79.8	74.9	80.7
w/o $L_{\text{syn-off}}$	46.88	84.78	84.76	0.721	0.791	0.709	0.657	0.724	79.4	74.2	80.2	55.0	85.38	85.30	0.536	0.767	0.696	0.644	0.710	80.1	74.8	81.0
w/o $L_{\text{syn-res}}$	46.43	84.32	84.30	0.731	0.786	0.716	0.682	0.725	80.0	76.0	80.4	54.6	84.92	84.84	0.545	0.762	0.704	0.664	0.710	80.8	76.8	81.4
w/o masked views	46.23	84.12	84.08	0.737	0.782	0.690	0.646	0.695	77.4	72.2	78.2	54.3	84.60	84.52	0.550	0.756	0.679	0.631	0.683	77.8	72.6	78.7
Expert collapse: single expert	45.83	83.62	83.58	0.744	0.776	0.672	0.631	0.679	76.6	71.2	77.4	53.8	84.00	83.92	0.560	0.748	0.660	0.616	0.665	76.8	70.8	77.8

Table 3: Ablations on MOSI and MOSEI. Clean metrics: ACC7 (A7), ACC2 (A2), F1, MAE, and Corr (C). Robustness is summarized by normalized AUC under Noise (N), Missingness (M), and Conflict (Cf), reported for Corr [N(C), M(C), Cf(C)] and ACC2 [N(A2), M(A2), Cf(A2)]. Stress-test protocols, severity grids, and the normalized-AUC computation are provided in Appendix C.

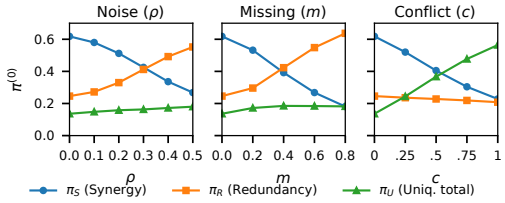

Figure 4: Router dynamics vs. corruption. Average initial routing weights $\mathbb{E}[\pi^{(0)}]$ under Noise (ρ), Missingness (m), and Cross-modal Conflict (c).

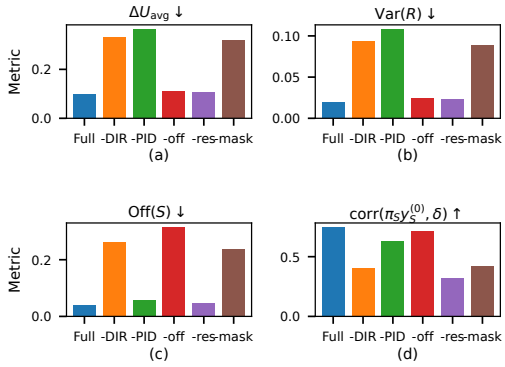

Figure 5: DIR semantic checks across variants. Lower is better for ΔU_{avg}, $\text{Var}(R)$, and $\text{Off}(S)$; higher is better for $\text{corr}(\pi_S^{(0)} y_S^{(0)}, \delta)$.

the target modality is present, (ii) *Redundancy stability* $\text{Var}(R)$ across the four views, (iii) *Synergy off-under-masking* $\text{Off}(S)$ on masked views, and (iv) *Synergy-as-residual* $\text{corr}(\pi_S^{(0)} y_S^{(0)}, \delta)$, where δ is the residual target from §3.4. Formal definitions (including view sets, aggregation, and the correlation computation) are given in Appendix C.4. The full model shows the intended signature: low ΔU, low $\text{Var}(R)$, near-zero $\text{Off}(S)$, and the strongest residual correlation. Ablations yield targeted degradations: removing L_{PID} primarily increases ΔU and $\text{Var}(R)$, indicating that uniqueness and redundancy lose their PID-inspired invariances; removing $L_{\text{syn-off}}$ sharply increases $\text{Off}(S)$ while leaving the other diagnostics comparatively close; and removing $L_{\text{syn-res}}$ most strongly reduces $\text{corr}(\pi_S y_S^{(0)}, \delta)$, consistent with losing the residual-fit constraint on synergy. Disabling DIR entirely (and training without masked views) degrades multiple metrics simultaneously, confirming that view perturbations and DIR are both required to prevent expert-role collapse.

Takeaway. Together, Figure 4 and Figure 5 support a closed-loop mechanism: the router performs sample-wise strategy switching under corruption, and DIR keeps the routed experts behaviorally disentangled so that these switches remain interpretable and robust.

5 Conclusion

In this paper, we studied robustness in multimodal sentiment analysis under sample-dependent modality unreliability and cross-modal disagreement. We proposed RIDER-MoE, a reliability-aware mixture-of-experts model that routes each input between synergy, redundancy, and modality-specific experts using uncertainty and consensus signals, and we introduced masked-view Disentangled Interaction Regularization (DIR) to encourage stable expert roles. Experiments on MOSI and MOSEI show that RIDER-MoE maintains competitive clean performance while improving robustness under noise, missing-modality, and conflict stress tests. Future work includes evaluating under more realistic distribution shifts and extending reliability-aware routing to additional modalities and tasks.

Limitations

Our empirical evaluation is restricted to two curated benchmarks (CMU-MOSI and CMU-MOSEI), and robustness is primarily assessed through three controlled, synthetic test-time stressors: within-modality Noise (ρ), Missingness (m), and Cross-modal Conflict (c). While these protocols yield reproducible performance–severity curves and AUC summaries, they may not faithfully capture real-world distribution shifts such as non-stationary background noise, temporal misalignment, correlated sensor failures, or other in-the-wild corruption patterns; thus, external validity beyond these benchmarks remains unverified.

Methodologically, masked-view training approximates "removing a modality" via a representation-level intervention (replacing an encoder output with a masking vector). This approximation relies on the assumption that the masking vector is label-agnostic and may differ from true absence, out-of-distribution inputs, or missing-not-at-random mechanisms. Likewise, our conflict construction breaks cross-modal agreement through controlled modality mismatch, which may not represent all natural forms of semantic disagreement.

Our PID framing is a motivated surrogate: we do not estimate mutual information, do not commit to a redundancy definition, and our Uniqueness/Redundancy/Synergy experts form a coarse aggregation over finer-grained PID atoms. Disentangled Interaction Regularization (DIR) imposes soft, weighted constraints that encourage (but do not guarantee) disentanglement and expert role specialization.

The reliability-aware router depends on discretized unimodal probes and uncertainty / agreement signals (e.g., entropy- and divergence-based consensus), which could be miscalibrated under sarcasm-like disagreement, severe corruption, or domain shift and may require recalibration. Finally, although inference is single-pass, training incurs additional computation and hyperparameter sensitivity due to multi-view objectives and regularization. We do not systematically analyze fairness/bias, privacy, or downstream deployment risks; we discuss these considerations in the Ethics Statement.

Ethics Statement

This work studies multimodal sentiment analysis from text, audio, and vision. Inferring affective states from human communication can be sensitive: the same technology may be used for beneficial applications (e.g., improving human–computer interaction) but also for harmful or high-stakes uses (e.g., surveillance, screening in hiring/education, or other automated decision-making about individuals). We therefore emphasize that our contributions are intended for research on robustness and reliability modeling, and we caution against deploying sentiment inference systems as the sole basis for consequential decisions without human oversight, clear user notice/consent, and rigorous risk assessment.

Data and privacy. The benchmarks used in this paper contain human-generated online video content with audio and visual signals. Such modalities can reveal identity-related or biometric cues (e.g., voice and facial characteristics). Even when datasets are publicly released for research, real-world deployment should follow data-minimization principles, appropriate access controls, and privacy-preserving handling (e.g., on-device processing, avoiding retention of raw audio/video when not necessary).

Bias and fairness. Performance may vary across demographic groups, languages/accents, and expression styles (including sarcasm or culturally specific affect). Our experiments do not provide a systematic audit of group fairness or representativeness. We recommend that future work evaluate subgroup performance and calibration, and consider bias mitigation or uncertainty-aware abstention when models are applied beyond the benchmark domain.

Deployment considerations. While we explicitly target reliability under controlled noise, missingness, and modality conflict, these stressors are synthetic and may not cover the full range of failures in the wild. Practical deployment should include ongoing monitoring, recalibration, and fallbacks (e.g., abstaining or deferring to robust unimodal pathways) when uncertainty or disagreement signals indicate low reliability.

References

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol-*

ume 1: Long Papers), pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10.

K. Chen, L. Xu, and H. Chi. 1999. Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*, 12(9):1229–1252.

Kezhou Chen, Shuo Wang, Huixia Ben, Shengeng Tang, and Yanbin Hao. 2025. Mixture of multimodal adapters for sentiment analysis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1822–1833, Albuquerque, New Mexico. Association for Computational Linguistics.

Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. 2023. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17300–17311.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep — a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. 2025. Emoe: Modality-specific enhanced dynamic emotion experts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14314–14324.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. *Preprint*, arXiv:2005.03545.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1309–1317.

Qian Li, Lixin Su, Jiashu Zhao, Long Xia, Hengyi Cai, Suqi Cheng, Hengzhu Tang, Junfeng Wang, and Dawei Yin. 2024. Text-video retrieval via variational multi-modal hypergraph networks. *Preprint*, arXiv:2401.03177.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. *Preprint*, arXiv:2303.13802.

Zuocheng Li and Lishuang Li. 2025. t-HNE: A text-guided hierarchical noise eliminator for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2834–2844, Abu Dhabi, UAE. Association for Computational Linguistics.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Russ R Salakhutdinov, and Louis-Philippe Morency. 2023. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pages 27351–27393. Curran Associates, Inc.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, page 169–176, New York, NY, USA. Association for Computing Machinery.

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

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *Preprint*, arXiv:2005.00357.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2023. Beneath the tip of

the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*, 14(1):108–132.

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

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019b. Multimodal transformer for unaligned multimodal language sequences. *Preprint*, arXiv:1906.00295.

Paul L. Williams and Randall D. Beer. 2010. Non-negative decomposition of multivariate information. *Preprint*, arXiv:1004.2515.

Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. 2023. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *J. Mach. Learn. Res.*, 24(1).

Han Wu, Yanming Sun, Yunhe Yang, and Derek F. Wong. 2025. Beyond simple fusion: Adaptive gated fusion for robust multimodal sentiment analysis. *Preprint*, arXiv:2510.01677.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L. Ballard, Tianlong Chen, and Qi Long. 2025. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. *Preprint*, arXiv:2505.19190.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Preprint*, arXiv:2102.04830.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *Preprint*, arXiv:1606.06259.

Xueliang Zhao, Mingyang Wang, Yingchun Tan, and Xianjie Wang. 2024. Tgmoe: A text guided mixture-of-experts model for multimodal sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 15(8).

Miao Zhou, Lina Yang, Thomas Wu, Dongnan Yang, and Xinru Zhang. 2025. Dual-path dynamic fusion with learnable query for multimodal sentiment analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11366–11376, Suzhou, China. Association for Computational Linguistics.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22123–22138, Vienna, Austria. Association for Computational Linguistics.

A PID-Consistent Interpretation of U/R/S Experts and DIR

A.1 PID background and scope of the theoretical claim

Let (T, A, V) denote random variables corresponding to the three modalities (Text/Audio/Vision), and let Y denote the sentiment label. Partial Information Decomposition (PID) (Williams and Beer, 2010; Wollstadt et al., 2023) is an information-theoretic framework that decomposes the total mutual information

$$I(Y; T, A, V) \tag{4}$$

into *information atoms* that separate *redundant* (shared), *unique* (source-specific), and *synergistic* (interaction-only) contributions. Formally, PID specifies a redundancy function (often denoted $I_\cap$) that determines which information about Y is shared among sources; once $I_\cap$ is fixed, the remaining atoms (including unique and synergy terms) are determined by the PID lattice construction (Williams and Beer, 2010; Wollstadt et al., 2023). As a consequence, PID *is not a single quantity*: the numerical values of "redundancy/uniqueness/synergy" depend on the chosen redundancy definition and estimator.

What we do (and do not) do. RIDER-MoE does *not* compute PID values (no explicit mutual-information estimation, no explicit choice of $I_\cap$). Instead, we use PID as a *theoretical motivation* for designing a trainable architecture and a set of masked-view constraints (DIR) that act as an *operational surrogate* for the core PID semantics: (i) uniqueness should not require other modalities, (ii) redundancy should persist under single-modality removal, and (iii) synergy should require all modalities and behave as an interaction-only correction. This "PID-consistent surrogate" viewpoint

is also aligned with recent uses of information-decomposition ideas for interpreting multimodal interactions.

A.2 RIDER-MoE notation and masked views as interventions

We follow the main text notation. For each modality $m \in \mathcal{M} = \{T, A, V\}$, a unimodal encoder produces $h_m = g_m(x_m)$. We define four *views* indexed by

$$p \in \{0, -T, -A, -V\}, \tag{5}$$

where $p = 0$ denotes the full tri-modal view and $p = -m$ denotes the masked view obtained by replacing h_m with a fixed masking vector MASK_m (Appendix B.2). For each view,

$$\begin{aligned} z^{(p)} &= f_{\text{fuse}}\Big(h_T^{(p)},\, h_A^{(p)},\, h_V^{(p)}\Big), \\ y_e^{(p)} &= f_e\Big(z^{(p)}\Big), \qquad \forall\, e \in \mathcal{E}. \end{aligned} \tag{6}$$

with expert set $\mathcal{E} = \{U_T, U_A, U_V, R, S\}$ as in §3.1. Routing weights are computed only on the full view,

$$\pi^{(0)} \in \mathbb{R}^{|\mathcal{E}|}, \quad \sum_{e\in\mathcal{E}} \pi_e^{(0)} = 1, \;\; \pi_e^{(0)} \geq 0, \tag{7}$$

and the full-view prediction is $\hat{y}^{(0)} = \sum_{e\in\mathcal{E}} \pi_e^{(0)} y_e^{(0)}$ (Eq. ((1))).

Masked views as operational "removal" interventions. The masking operation is an implementation of a controlled intervention that removes the information content of one modality while keeping the remaining modalities fixed at the representation level. Concretely, $p = -m$ corresponds to the intervention $h_m \leftarrow \mathrm{MASK}_m$. Our DIR arguments below rely on the mild assumption that MASK_m is *label-agnostic* (it does not itself carry information about Y), so that comparing $y_e^{(0)}$ and $y_e^{(-m)}$ probes sensitivity to the presence of modality m rather than injecting new label information.

A.3 DIR losses as PID-semantic surrogate constraints

DIR consists of the uniqueness and redundancy invariance terms ($L_{\text{PID}} = L_{\text{uni}} + L_{\text{red}}$) and synergy regularization ($L_{\text{syn}} = L_{\text{syn-off}} + L_{\text{syn-res}}$), defined in Appendix B.3. For completeness, we restate the key structures: **Uniqueness invariance.**

$$\begin{aligned} L_{\text{uni},m} &= \frac{1}{|P_m| - 1} \sum_{p\in P_m\setminus\{0\}} \big(y_{U_m}^{(0)} - y_{U_m}^{(p)}\big)^2, \\ L_{\text{uni}} &= \sum_{m\in\mathcal{M}} r_m\, L_{\text{uni},m}. \end{aligned} \tag{8}$$

Redundancy stability.

$$\begin{aligned} \bar{y}_R &= \frac{1}{4} \sum_{p\in\{0,-T,-A,-V\}} y_R^{(p)}, \\ L_{\text{red}} &= \bar{r} \cdot \frac{1}{4} \sum_{p\in\{0,-T,-A,-V\}} \big(y_R^{(p)} - \bar{y}_R\big)^2. \end{aligned} \tag{9}$$

Synergy off under masking.

$$L_{\text{syn-off}} = \bar{r}^{\,3} \cdot \frac{1}{3} \sum_{p\in\{-T,-A,-V\}} \big(y_S^{(p)}\big)^2. \tag{10}$$

Synergy as residual-fit.

$$\begin{aligned} \hat{y}_{\setminus S}^{(0)} &= \sum_{e\in\mathcal{E}\setminus\{S\}} \pi_e^{(0)} y_e^{(0)}, \\ \delta &= y - \mathrm{sg}\Big(\hat{y}_{\setminus S}^{(0)}\Big), \\ L_{\text{syn-res}} &= \bar{r}^{\,3} \big(\pi_S^{(0)} y_S^{(0)} - \delta\big)^2. \end{aligned} \tag{11}$$

Here $P_T = \{0, -A, -V\}$, $P_A = \{0, -T, -V\}$, $P_V = \{0, -T, -A\}$ (Eq. ((25))), and $\mathrm{sg}(\cdot)$ is stop-gradient.

Below we state a PID-consistent interpretation of each DIR component, as an *operational* surrogate for the corresponding PID semantics.

A.3.1 Uniqueness invariance as a proxy for modality-specificity

In PID terms, "unique" information in modality m refers to information about Y that can be obtained from m without requiring the other modalities (and that is not available from them, depending on $I_\cap$). Our uniqueness invariance term targets the *necessary* (but not sufficient) operational property: the uniqueness expert should be insensitive to removing other modalities.

Proposition 1 (Uniqueness invariance discourages cross-modal leakage). Fix a modality $m \in \mathcal{M}$ and define the composite mapping

$$G_{U_m}(h_T, h_A, h_V) \;:=\; f_{U_m}\Big(f_{\text{fuse}}(h_T, h_A, h_V)\Big). \tag{12}$$

Assume that masking $m' \neq m$ replaces $h_{m'}$ by a constant vector $\mathrm{MASK}_{m'}$. If $L_{\text{uni},m} = 0$ holds for

all examples (i.e., $y^{(0)}_{U_m} = y^{(p)}_{U_m}$ for all $p \in P_m \backslash \{0\}$), then G_{U_m} is invariant to each masked intervention on modalities $m' \neq m$ on the support of the data:

$$G_{U_m}(h_T, h_A, h_V) = G_{U_m}\left(h_T^{(-m')}, h_A^{(-m')}, h_V^{(-m')}\right) \quad \forall\, m' \in \mathcal{M} \setminus \{m\}. \tag{13}$$

and thus $y^{(0)}_{U_m}$ can be written as a function of h_m and constants (the masks) only, i.e., there exists ϕ_m such that

$$y^{(0)}_{U_m} \;=\; \phi_m(h_m) \quad \text{on the data support.} \tag{14}$$

Proof sketch. For $m = T$ (analogous for A, V), $L_{\text{uni},T} = 0$ implies

$$\begin{aligned} G_{U_T}(h_T, h_A, h_V) &= G_{U_T}(h_T, \text{MASK}_A, h_V) \\ &= G_{U_T}(h_T, h_A, \text{MASK}_V). \end{aligned} \tag{15}$$

The first equality states that changing h_A to a constant does not change the output, so G_{U_T} is insensitive to h_A under this intervention; the second equality analogously removes sensitivity to h_V. Therefore, on the data support, the output depends only on the remaining variable h_T (plus fixed constants), yielding a representation $\phi_T(h_T)$.

PID-semantic proxy. This proposition does *not* claim that U_m captures *only* PID-unique information (which would require a specific redundancy definition and information estimation). It formalizes the narrower, PID-consistent surrogate: the U_m expert is trained to be *modality-specific* in the sense of being invariant to removing other modalities, thereby discouraging cross-modal leakage into U_m.

A.3.2 Redundancy stability as a proxy for shared information

PID redundancy corresponds to information about Y that is available from multiple sources. Our redundancy stability term L_{red} enforces that R outputs remain stable under removing any single modality, which operationalizes "shared" information as *persistence under single-modality removal*.

Proposition 2 (Variance penalty suppresses modality-specific sensitivity). Assume a simple shared+specific factorization of representations: for each $m \in \mathcal{M}$,

$$h_m \;=\; \big[c\,;\, u_m\big], \tag{16}$$

where c is a shared factor (correlated across modalities and predictive of Y) and u_m is a modality-specific factor (not shared with other modalities). Consider a setting where the redundancy output on a view is locally linear in the three modality representations,

$$y_R^{(p)} \approx b + \sum_{m \in \mathcal{M}} w_m^\top h_m^{(p)}, \tag{17}$$

and masking sets $h_m^{(-m)} = \text{MASK}_m$ with fixed $\text{MASK}_m = [c_0; u_0]$. Then, under mild moment assumptions (e.g., $\mathbb{E}[u_m] = 0$ and the u_m are not perfectly predictable from the other modalities), minimizing $\mathbb{E}[L_{\text{red}}]$ discourages placing weight on modality-specific directions, i.e., it drives w_m to have small components on u_m-subspaces, making $y_R^{(p)}$ primarily a function of the shared factor c (which remains recoverable from the unmasked modalities).

Proof sketch. Across the four views, the only changes induced by masking are replacements of one modality vector by a constant. In the local linear approximation, the view-to-view variability of $y_R^{(p)}$ is controlled by the magnitudes of $w_m^\top(h_m - \text{MASK}_m)$. If w_m has nontrivial projection onto u_m directions that fluctuate across samples and are not reproducible from the other modalities, then masking modality m produces view-dependent deviations that increase the variance term in Eq. (9). Reducing this variance thus pushes w_m away from those modality-specific directions, leaving the stable (shared) signal as the remaining low-variance predictor.

PID-semantic proxy. The proposition formalizes a PID-consistent surrogate notion of redundancy: R is encouraged to represent components that are stable under removing any one modality, which is a natural operational proxy for "shared" information across modalities.

A.3.3 Synergy off-under-masking as a proxy for joint-only information

In PID, synergy corresponds to information about Y that emerges only when sources are observed jointly. Our $L_{\text{syn-off}}$ implements the operational constraint that the synergy expert should be silent whenever any modality is removed.

Proposition 3 (Off-under-masking enforces joint-only activation). If $L_{\text{syn-off}} = 0$ for all

examples, then for each masked view $p \in \{-T, -A, -V\}$,

$$y_S^{(p)} = 0. \tag{18}$$

Consequently, the synergy pathway cannot produce any expert-level signal under any single-modality removal intervention, and any nontrivial contribution of S must come from the full view $p = 0$.

Proof sketch. Eq. (10) is a sum of squared terms. Setting it to zero forces each squared term to be zero, hence $y_S^{(p)} = 0$ on the masked views.

PID-semantic proxy. This is a direct operationalization of the defining intuition behind synergy: it should vanish when a modality is missing, and thus can only reflect information accessible from the joint tri-modal context.

A.3.4 Synergy-as-residual as a proxy for interaction-only correction

The second synergy constraint is residual-fit on the full view. It encourages S to explain what is *not* already explained by the other experts, closely matching the intuition of "interaction-only" content.

Proposition 4 (Residual-fit makes synergy approximate the leftover error). Condition on the full-view representation $z^{(0)}$ and treat the non-synergy prediction $\hat{y}_{\setminus S}^{(0)}$ as fixed with respect to the gradients of $L_{\text{syn-res}}$ via $\mathrm{sg}(\cdot)$. Then minimizing the expected residual-fit term over the data distribution,

$$\min_S \mathbb{E}\Big[\big(\pi_S^{(0)} y_S^{(0)} - \delta\big)^2\Big], \tag{19}$$

drives the *routed synergy contribution* $\pi_S^{(0)} y_S^{(0)}$ toward the conditional least-squares predictor of the residual:

$$\pi_S^{(0)} y_S^{(0)} \;\approx\; \mathbb{E}\Big[\delta \mid z^{(0)}\Big] \;=\; \mathbb{E}\Big[y - \hat{y}_{\setminus S}^{(0)} \mid z^{(0)}\Big]. \tag{20}$$

Thus, S is trained to model the portion of Y that remains unexplained by U_T, U_A, U_V, and R under the current routing policy.

Derivation sketch. Under squared loss, the minimizer of $\mathbb{E}[(g(z^{(0)}) - \delta)^2]$ over measurable functions g is $g^\star(z^{(0)}) = \mathbb{E}[\delta \mid z^{(0)}]$. Here $g(z^{(0)})$ is instantiated by the routed synergy contribution $\pi_S^{(0)} y_S^{(0)}$. The stop-gradient in Eq. (11) prevents this objective from changing $\hat{y}_{\setminus S}^{(0)}$ through gradient flow, making the residual target δ act as a fixed regression target for the synergy pathway during this update.

PID-semantic proxy. Residual-fit does not compute a PID synergy atom. It enforces a PID-consistent *operational* surrogate: synergy is trained as an interaction-only correction beyond what can already be captured by the (masked-invariant) uniqueness and redundancy pathways.

A.4 Assumptions and limitations of the PID surrogate (mandatory)

The bridge above is intentionally framed as a *PID-consistent surrogate*, not as an equality to PID quantities. Key limitations are:

- **Supervised objective vs. information quantities.** DIR optimizes prediction-space constraints on scalar expert outputs (Eqs. (8)–(11)) alongside the task loss (Eq. ((2))); it does *not* estimate mutual information, and thus does not yield numerical PID decompositions.
- **No fixed redundancy definition.** PID depends on a redundancy function $I_\cap$ (Williams and Beer, 2010; Wollstadt et al., 2023). RIDER-MoE does not choose or estimate $I_\cap$, so the expert outputs cannot be interpreted as *the* PID atoms under any specific PID instantiation.
- **Masked views approximate "removal".** Masking operates at the representation level and is an approximation to removing a modality. If MASK_m is imperfect (e.g., carries unintended cues), the intervention interpretation weakens.
- **Coarse PID categories.** Full tri-source PID contains multiple redundancy and synergy atoms across subsets of sources; our U/R/S split is a coarser grouping into modality-specific, robust shared, and full-view interaction-only pathways.
- **Constraints are soft.** In practice, losses are weighted (Eq. ((3))) and optimized approximately; we therefore claim *encouragement* toward PID semantics, not a strict guarantee of disentanglement.

Practical Takeaway for the Main Text Claim. When the main text says "PID-inspired/grounded," the intended claim is: *PID provides principled semantics for uniqueness/redundancy/synergy, and DIR operationalizes these semantics via masked-view invariances and residual constraints (Appendix A); we do not compute PID values or claim strict equality to any particular PID decomposition.*

B Additional Details for the Method

This appendix provides the full mathematical definitions and algorithmic details underlying the Method in Section 3. We retain the notation from the main text.

B.1 Reliability-Aware Router Details

We expand the reliability-aware routing scheme introduced in Section 3.2.

B.1.1 Unimodal sentiment probes

For each modality $m \in \mathcal{M}$, we attach a small classification head to the encoder output h_m:

$$\ell_m^{\text{cls}} = W_m^{\text{cls}} h_m + b_m^{\text{cls}} \in \mathbb{R}^K, \tag{21}$$

where K is the number of discrete sentiment levels used to discretize the continuous label (e.g., $K = 7$ for scores in $\{-3, \dots, 3\}$). A softmax yields the unimodal sentiment distribution

$$p_m(k) = \frac{\exp(\ell_{m,k}^{\text{cls}})}{\sum_{k'=1}^{K} \exp(\ell_{m,k'}^{\text{cls}})}, \quad k = 1, \dots, K. \tag{22}$$

Let $\tilde{y} \in \{1, \dots, K\}$ denote the discretized label. The probe for modality m is trained with cross-entropy

$$\mathcal{L}_{\text{cls}}^{(m)} = \text{CE}(p_m, \tilde{y}). \tag{23}$$

B.1.2 Entropy-based modality reliability

From each unimodal distribution p_m, we compute the entropy

$$H_m = -\sum_{k=1}^{K} p_m(k) \, \log\big(p_m(k) + \varepsilon\big), \quad m \in \mathcal{M}, \tag{24}$$

with a small $\varepsilon > 0$ for numerical stability. Lower entropy corresponds to a sharper, more confident unimodal prediction. We convert entropies into a relative reliability distribution

$$\text{Rel}_m = \frac{\exp(-H_m)}{\sum_{m' \in \mathcal{M}} \exp(-H_{m'})}, \quad m \in \mathcal{M}. \tag{25}$$

B.1.3 Modality importance from fused context

To capture how useful each modality appears from the perspective of the fused representation, we predict modality importance directly from $z^{(0)}$. A linear layer followed by softmax yields

$$u = W_{\text{imp}} z^{(0)} + b_{\text{imp}} \in \mathbb{R}^3, \tag{26}$$

$$\text{Imp}_m = \frac{\exp(u_m)}{\sum_{m' \in \mathcal{M}} \exp(u_{m'})}, \quad m \in \mathcal{M}. \tag{27}$$

B.1.4 Modality priors and reliability weights

Reliability and importance are blended into a modality-level prior. Let $\alpha \in [0, 1]$ control their trade-off. We first form

$$\tilde{r}_m = \alpha \, \text{Rel}_m + (1 - \alpha) \, \text{Imp}_m, \quad m \in \mathcal{M}, \tag{28}$$

then normalize

$$r_m = \frac{\tilde{r}_m}{\sum_{m' \in \mathcal{M}} \tilde{r}_{m'}}, \quad \bar{r} = \frac{1}{3}\big(r_T + r_A + r_V\big). \tag{29}$$

The r_m act as reliability-based weights for modality-specific regularization, and $\bar{r}$ summarizes joint reliability of all three modalities.

B.1.5 Consensus score from unimodal predictions

We quantify agreement between modalities by comparing each p_m to the average distribution

$$\bar{p} = \frac{1}{3}\big(p_T + p_A + p_V\big). \tag{30}$$

For each modality,

$$D_m = \text{KL}(p_m \,\|\, \bar{p}) = \sum_{k=1}^{K} p_m(k) \, \log \frac{p_m(k) + \varepsilon}{\bar{p}(k) + \varepsilon}, \tag{31}$$

and the mean disagreement is

$$\text{Disagree} = \frac{1}{3}\big(D_T + D_A + D_V\big). \tag{32}$$

A scalar consensus score is then defined as

$$\text{cons\_score} = \exp\big(-\gamma \cdot \text{Disagree}\big), \quad \gamma > 0. \tag{33}$$

High cons_score indicates that unimodal probes make similar predictions; low values reflect cross-modal conflict, where **uniqueness** (and, when appropriate, redundancy) experts should be emphasized while the **synergy** expert is down-weighted to avoid propagating inconsistent signals.

B.1.6 Expert priors and log-prior injection

We summarize modality-level information into expert-level priors. A feature vector

$$f = \begin{bmatrix} \mathrm{Prior}_T^{\mathrm{mod}} \\ \mathrm{Prior}_A^{\mathrm{mod}} \\ \mathrm{Prior}_V^{\mathrm{mod}} \\ \mathrm{cons\_score} \end{bmatrix} \in \mathbb{R}^4 \tag{34}$$

is constructed from the modality priors $\mathrm{Prior}_m^{\mathrm{mod}} := r_m$ and the consensus score. A linear map produces expert logits

$$u^{\mathrm{exp}} = W_{\mathrm{prior}} f + b_{\mathrm{prior}} \in \mathbb{R}^5, \tag{35}$$

which are converted into an expert prior distribution

$$\mathrm{Prior}_e^{\mathrm{exp}} = \frac{\exp(u_e^{\mathrm{exp}})}{\sum_{e' \in \mathcal{E}} \exp(u_{e'}^{\mathrm{exp}})}, \quad e \in \mathcal{E}. \tag{36}$$

In parallel, a parametric router network computes base logits from the fused representation:

$$z_{\mathrm{router}} = f_{\mathrm{router}}\left(z^{(0)}\right) \in \mathbb{R}^5. \tag{37}$$

We inject the expert prior as a log-bias with strength $\lambda_p \geq 0$:

$$\tilde{z}_{\mathrm{router},e} = z_{\mathrm{router},e} + \lambda_p \log\left(\mathrm{Prior}_e^{\mathrm{exp}} + \varepsilon\right), \quad e \in \mathcal{E}, \tag{38}$$

and obtain the final routing probabilities by softmax

$$\pi_e^{(0)} = \frac{\exp\left(\tilde{z}_{\mathrm{router},e}\right)}{\sum_{e' \in \mathcal{E}} \exp\left(\tilde{z}_{\mathrm{router},e'}\right)}. \tag{39}$$

Thus the router remains trained end-to-end through $\mathcal{L}_{\mathrm{task}}$, while being softly guided by reliability- and consensus-aware expert priors.

B.2 Interaction Experts and Masked Views

We provide the explicit construction of masked views used in Section 3.3. After computing (h_T, h_A, h_V), we construct four views indexed by p:

- Full view ($p = 0$):

$$(h_T^{(0)}, h_A^{(0)}, h_V^{(0)}) = (h_T, h_A, h_V).$$

- Text-masked view ($p = -T$):

$$(h_T^{(-T)}, h_A^{(-T)}, h_V^{(-T)}) = (\mathrm{MASK}_T,\ h_A,\ h_V).$$

- Audio-masked view ($p = -A$):

$$(h_T^{(-A)}, h_A^{(-A)}, h_V^{(-A)}) = (h_T,\ \mathrm{MASK}_A,\ h_V).$$

- Vision-masked view ($p = -V$):

$$(h_T^{(-V)}, h_A^{(-V)}, h_V^{(-V)}) = (h_T,\ h_A,\ \mathrm{MASK}_V).$$

Here MASK_m is a fixed or randomly sampled vector in the same space as h_m that effectively removes modality m.

For each view $p \in \{0, -T, -A, -V\}$, we reuse the same fusion module and experts:

$$z^{(p)} = f_{\mathrm{fuse}}\big(h_T^{(p)}, h_A^{(p)}, h_V^{(p)}\big), \tag{40}$$

$$y_e^{(p)} = f_e\left(z^{(p)}\right), \quad e \in \mathcal{E}. \tag{41}$$

All routing-related quantities are computed once from the full view $(h_T, h_A, h_V, z^{(0)})$, and the resulting $\pi^{(0)}$ is reused for all views whenever routing weights are needed in a loss.

B.3 Disentangled Interaction Regularization Details

We now provide the exact formulas for the DIR losses described in Section 3.4.

B.3.1 PID-inspired invariance for uniqueness and redundancy

For each modality $m \in \{T, A, V\}$, define the set of views where m is present:

$$\begin{aligned} P_T &= \{0, -A, -V\}, \\ P_A &= \{0, -T, -V\}, \\ P_V &= \{0, -T, -A\}. \end{aligned} \tag{42}$$

We want the corresponding uniqueness expert U_m to give consistent predictions across these views. For modality m,

$$\mathcal{L}_{\mathrm{uni},m} = \frac{1}{|P_m| - 1} \sum_{p \in P_m \setminus \{0\}} \big(y_{U_m}^{(0)} - y_{U_m}^{(p)}\big)^2. \tag{43}$$

The global uniqueness loss, weighted by modality reliability, is

$$\mathcal{L}_{\mathrm{uni}} = \sum_{m \in \mathcal{M}} r_m \, \mathcal{L}_{\mathrm{uni},m}. \tag{44}$$

The redundancy expert R should capture information that persists when any single modality is perturbed. We therefore encourage its output to be stable across all four views. Let

$$\bar{y}_R = \frac{1}{4} \sum_{p \in \{0, -T, -A, -V\}} y_R^{(p)}, \tag{45}$$

and define

$$\mathcal{L}_{\text{red}} = \bar{r} \cdot \frac{1}{4} \sum_{p \in \{0,-T,-A,-V\}} \left(y_R^{(p)} - \bar{y}_R\right)^2. \quad (46)$$

This variance penalty discourages R from encoding modality-specific fluctuations; the only information it can represent consistently is that which survives dropping any single modality.

The PID-inspired invariance loss is

$$\mathcal{L}_{\text{PID}} = \mathcal{L}_{\text{uni}} + \mathcal{L}_{\text{red}}. \quad (47)$$

B.3.2 Synergy regularization

The second DIR component, $\mathcal{L}_{\text{syn}}$, enforces a synergy interpretation.

Synergy off under masking. When any modality is masked, synergy should vanish. We penalize non-zero synergy outputs on masked views:

$$\mathcal{L}_{\text{syn-off}} = \bar{r}^3 \cdot \frac{1}{3} \sum_{p \in \{-T,-A,-V\}} \left(y_S^{(p)}\right)^2. \quad (48)$$

The factor $\bar{r}^3$ emphasizes examples where all three modalities are simultaneously reliable; when modalities are noisy, synergy is not strongly constrained.

Residual fit on the full view. On the full view, the synergy contribution should explain whatever part of the label cannot be captured by the uniqueness and redundancy experts. Let

$$\hat{y}_{\setminus S}^{(0)} = \sum_{e \in \mathcal{E} \setminus \{S\}} \pi_e^{(0)} \, y_e^{(0)}, \quad (49)$$

and define a residual target with stop-gradient

$$\delta = y - \text{sg}\left(\hat{y}_{\setminus S}^{(0)}\right), \quad (50)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. The routed synergy contribution is encouraged to match this residual:

$$\mathcal{L}_{\text{syn-res}} = \bar{r}^3 \left(\pi_S^{(0)} y_S^{(0)} - \delta\right)^2. \quad (51)$$

Because gradients are stopped through $\hat{y}_{\setminus S}^{(0)}$, this term updates only the synergy expert S and its routing weight $\pi_S^{(0)}$; the other experts are optimized solely via $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{PID}}$.

Total synergy loss. The synergy loss is

$$\mathcal{L}_{\text{syn}} = \mathcal{L}_{\text{syn-off}} + \mathcal{L}_{\text{syn-res}}. \quad (52)$$

Together, $\mathcal{L}_{\text{syn-off}}$ and $\mathcal{L}_{\text{syn-res}}$ enforce that synergy is silent when any modality is missing and acts as an interaction-only correction on the full view.

Algorithm 1 Training and inference for reliability-aware interaction MoE

Input: Mini-batch $\{(x_T, x_A, x_V, y)\}$, model parameters.
Training: For each mini-batch:
  Encode each modality: $h_m \leftarrow g_m(x_m)$ for $m \in \{T, A, V\}$.
  Compute unimodal probe logits ℓ_m^{cls}, distributions p_m, entropies H_m, and losses $\mathcal{L}_{\text{cls}}^{(m)}$.
  Form full fused representation $z^{(0)} \leftarrow f_{\text{fuse}}(h_T, h_A, h_V)$.
  From $z^{(0)}$ and $\{p_m\}$, compute modality priors r_m, consensus score, expert priors $\text{Prior}^{\text{exp}}$, router logits, and routing probabilities $\pi^{(0)}$.
  Construct masked encoder outputs $(h_T^{(p)}, h_A^{(p)}, h_V^{(p)})$ for $p \in \{-T, -A, -V\}$ and set $(h_T^{(0)}, h_A^{(0)}, h_V^{(0)}) = (h_T, h_A, h_V)$.
  For each view $p \in \{0, -T, -A, -V\}$:
    $z^{(p)} \leftarrow f_{\text{fuse}}(h_T^{(p)}, h_A^{(p)}, h_V^{(p)})$;
    $y_e^{(p)} \leftarrow f_e(z^{(p)})$ for all $e \in \mathcal{E}$.
  Compute $\mathcal{L}_{\text{task}}$ using Eq. (2) and $\hat{y}^{(0)}$ from Eq. (1).
  Compute $\mathcal{L}_{\text{PID}}$ and $\mathcal{L}_{\text{syn}}$ from expert outputs across views, always routing with $\pi^{(0)}$.
  Form total loss $\mathcal{L}$ using Eq. (3) and update all parameters by back-propagation.
Inference: For a test input (x_T, x_A, x_V):
  Encode (h_T, h_A, h_V) and compute $z^{(0)}$.
  Optionally compute probes and reliability signals to obtain $\pi^{(0)}$.
  Compute expert outputs $y_e^{(0)}$ and final prediction $\hat{y}^{(0)} = \sum_e \pi_e^{(0)} y_e^{(0)}$.

B.4 Training Algorithm

Algorithm 1 summarizes the training and inference procedure corresponding to Section 3. It uses the task loss in Eq. (2), the prediction in Eq. (1), and the regularization terms defined above.

C Experiments

C.1 Datasets

CMU-MOSI consists of 2,199 monologue video samples, with 1,284 for training, 229 for validation, and 686 for testing. Acoustic and visual features are sampled at 12.5 Hz and 15 Hz, respectively. **CMU-MOSEI** includes 22,856 YouTube movie review clips, with 16,326 for training, 1,871 for validation, and 4,659 for testing. Acoustic and

visual features are sampled at 20 Hz and 15 Hz. Both datasets have sentiment labels ranging from -3 (highly negative) to 3 (highly positive).

C.2 Robustness stress-test protocols

We evaluate robustness using performance–severity curves under three test-time stressors: Noise (ρ), Missingness (m), and Cross-modal Conflict (c). Unless otherwise stated, we apply the stressor to the test set only (training remains unchanged) and evaluate all methods on the *same* corrupted inputs for fair comparison (i.e., we reuse the same corrupted test set at each severity across methods).

Severity grids. We sweep $\rho \in \{0, 0.1, \ldots, 0.5\}$, $m \in \{0, 0.2, \ldots, 0.8\}$, and $c \in \{0, 0.25, 0.5, 0.75, 1.0\}$, matching § 4.2.

Noise (ρ): within-modality corruption. Noise simulates sample-dependent quality degradation while preserving modality availability. For text, we randomly mask a fraction ρ of tokens (replacing them with a dedicated `[MASK]` token or an embedding-level mask used by the text encoder). For audio and vision, we perturb continuous features by (i) additive zero-mean Gaussian noise whose standard deviation scales with ρ and (ii) temporal dropout that zeros a fraction ρ of timesteps. This produces a monotonic quality degradation within each modality without explicitly removing modalities.

Missingness (m): modality dropout. Missingness simulates sensor failures by removing one or more modalities at test time. For each test instance, each modality $m \in \{T, A, V\}$ is independently dropped with probability m (we resample if all three are dropped). Dropping a modality is implemented by replacing its encoder output h_m with the same masking vector MASK_m used to construct masked views in §3.3 (Appendix B.2/B.3), i.e., in the fused input we set the missing modality representation to MASK_m and keep the other modalities unchanged. All baselines receive the same masked inputs; if a baseline includes an explicit missing-modality mechanism, we apply its intended inference procedure on top of the masked inputs.

Cross-modal Conflict (c): mismatched modalities that break agreement. Conflict simulates semantic disagreement by breaking cross-modal consistency while keeping unimodal statistics realistic. With probability c, we replace one or more non-text modalities (audio and/or vision) of a test instance with that from a *different* test instance, while keeping the original text and label fixed. To increase the likelihood of semantic conflict, we preferentially swap from an instance with opposite sentiment polarity (large label distance) when available. This preserves realistic marginal distributions for each modality but disrupts cross-modal alignment/agreement, directly probing whether a model can avoid harmful synergistic fusion under mismatch.

C.3 Robustness summary metric: normalized AUC

For a metric $M(\cdot)$ evaluated at a set of severities $\{s_0 < s_1 < \cdots < s_K\}$ (e.g., $s = \rho, m, c$), we compute the (discrete) area under the performance–severity curve via the trapezoidal rule:

$$\mathrm{AUC}(M) = \sum_{k=1}^{K} (s_k - s_{k-1}) \cdot \frac{M(s_k) + M(s_{k-1})}{2}. \tag{53}$$

We report *normalized AUC* by dividing by the severity range:

$$\mathrm{nAUC}(M) = \frac{\mathrm{AUC}(M)}{s_K - s_0}, \tag{54}$$

which is equivalent to the average performance over the severity grid when using trapezoidal integration. In the main paper, robustness AUC is reported for higher-is-better metrics (Acc-2 and Corr).

C.4 DIR semantic-check metrics

DIR is trained using the full view $p = 0$ and three masked views $p \in \{-T, -A, -V\}$ (Appendix B.3), but inference uses only the full view. For analysis (Figures 5), we additionally compute masked-view expert outputs at test time as diagnostics, reusing the same encoders/fusion module. Let $y_e^{(p)}$ denote the scalar output of expert $e \in \{U_T, U_A, U_V, R, S\}$ on view p, and let $\pi^{(0)}$ be the routing weights computed from the full view.

(1) Uniqueness consistency: ΔU_{avg}. For each modality $m \in \{T, A, V\}$, define the set of views where m is present (as in Appendix B.3): $P_T = \{0, -A, -V\}$, $P_A = \{0, -T, -V\}$, $P_V = \{0, -T, -A\}$. We measure how stable the corresponding uniqueness expert is across those views:

$$\Delta U_m = \mathbb{E}_i \left[\frac{1}{|P_m| - 1} \sum_{p \in P_m \setminus \{0\}} \left| y_{U_m}^{(0)}(i) - y_{U_m}^{(p)}(i) \right| \right]. \tag{55}$$

and report $\Delta U_{\text{avg}} = \frac{1}{3}\sum_{m\in\{T,A,V\}} \Delta U_m$. Lower ΔU indicates less cross-modal leakage into the uniqueness experts.

(2) Redundancy stability: $\mathrm{Var}(R)$. We measure how invariant the redundancy expert is to removing any single modality:

$$\mathrm{Var}(R) = \mathbb{E}_i \left[\mathrm{Var}_{p\in\{0,-T,-A,-V\}} \left(y_R^{(p)}(i) \right) \right], \tag{56}$$

where Var is the sample variance over the four views. Lower is better.

(3) Synergy off-under-masking: $\mathrm{Off}(S)$. Synergy should be silent on masked views; we quantify unintended activation as

$$\mathrm{Off}(S) = \mathbb{E}_i \left[\frac{1}{3} \sum_{p\in\{-T,-A,-V\}} \left| y_S^{(p)}(i) \right| \right]. \tag{57}$$

Lower is better.

(4) Synergy-as-residual: $\mathrm{corr}(\pi_S^{(0)} y_S^{(0)}, \delta)$. On the full view, DIR constrains the routed synergy contribution to match the residual target $\delta = y - \mathrm{sg}(\hat{y}_{\setminus S}^{(0)})$ defined in §3.4 (Appendix B.3), where $\hat{y}_{\setminus S}^{(0)} = \sum_{e\neq S} \pi_e^{(0)} y_e^{(0)}$. As a test-time diagnostic, we compute δ with the same expression (the stop-gradient is irrelevant for evaluation), and report the Pearson correlation across test instances:

$$\mathrm{corr}\left(\pi_S^{(0)} y_S^{(0)},\, \delta \right) = \mathrm{PearsonCorr}\Big(\{\pi_S^{(0)}(i) y_S^{(0)}(i)\}_i,\, \{\delta(i)\}_i \Big) \tag{58}$$

Higher correlation indicates that synergy behaves as a residual correction rather than encoding spurious unimodal information.

C.5 Implementation details

For CMU-MOSI and CMU-MOSEI, we utilize 300-dimensional GloVe language features (Pennington et al., 2014) and 768-dimensional BERT-base-uncased hidden states (Devlin et al., 2019). Facet (Baltrušaitis et al., 2016) provides 35 facial action unit visual features, and COVAREP (Degottex et al., 2014) offers 74-dimensional acoustic features. The reported results use hyperparameters selected on the validation set, and we report the corresponding test performance under the same training conditions (no test-set tuning). Experiments are conducted on a PyTorch framework using an A100 GPU with 40GB memory, with a batch size of 32 and training for 60 epochs. For all experiments, we use the AdamW optimizer. On **CMU-MOSI**, we set the MoE hidden size to $d_{\text{model}} = 128$ and instantiate Transformer-based experts with $L_e = 4$ layers, $H_e = 4$ attention heads, feed-forward dimension $d_{\text{ff}} = 256$, and dropout rate 0.3. The probe classifier uses $K = 7$ classes, and the reliability-aware routing adopts $\alpha = 0.7$, $\gamma = 1.0$, and prior-injection weight $\lambda_p = 1.0$. The loss weights are $\lambda_{\text{cls}} = 0.05$, $\lambda_{\text{PID}} = 0.05$, and $\lambda_{\text{syn}} = 0.05$. We use a learning rate of 3×10^{-4} with weight decay 5×10^{-4}. On **CMU-MOSEI**, we increase the model capacity by setting $d_{\text{model}} = 512$ and using Transformer experts with $L_e = 8$ layers, $H_e = 8$ heads, $d_{\text{ff}} = 1024$, and dropout 0.3. The probe and router hyperparameters remain the same ($K = 7$, $\alpha = 0.7$, $\gamma = 1.0$, $\lambda_p = 1.0$), while the loss weights are $\lambda_{\text{cls}} = 0.1$, $\lambda_{\text{PID}} = 0.05$, and $\lambda_{\text{syn}} = 0.05$. We use a learning rate of 5×10^{-4} with weight decay 1×10^{-4}.

C.6 Hyperparameter Sensitivity

We sweep the key routing and regularization hyperparameters around the defaults used in all main experiments: prior-injection strength $\lambda_p = 1.0$ and DIR weights $\lambda_{\text{PID}} = \lambda_{\text{syn}} = 0.05$. We report the same clean metrics as Table 3 (Acc-7/Acc-2/F1/MAE/Corr) and robustness summaries as normalized AUC (nAUC; ↑) under Noise/Missing/Conflict stress tests.

As λ_p increases from 0, both clean and robust scores improve smoothly and then saturate close to the default. Removing the consensus term mainly affects Conflict robustness, consistent with Table 3.

Across the tested neighborhood of $(\lambda_{\text{PID}}, \lambda_{\text{syn}})$, robustness improves gradually from the w/o DIR baseline and peaks near the default. Ratio sweeps show the expected trade-off: emphasizing λ_{PID} helps Conflict more, while emphasizing λ_{syn} more strongly benefits Noise/Missing.

Takeaway. The model is most sensitive to disabling DIR entirely; moderate changes around the default loss weights yield small, smooth variations in both clean and robust metrics. Consensus features are particularly important for Conflict robustness, whereas masked-view training is necessary to realize the DIR benefits under distribution shift. Overall, the default setting is near a plateau in both routing and DIR sweeps, suggesting that the reported configuration is not finely tuned to a narrow hyperparameter choice.

Setting	A7↑	A2↑	F1↑	MAE↓	C↑	N(C)↑	M(C)↑	Cf(C)↑	N(A2)↑	M(A2)↑	Cf(A2)↑
CMU-MOSI											
$\lambda_p = 0^\dagger$	46.93	84.82	84.80	0.720	0.792	0.709	0.671	0.724	79.3	75.0	80.2
$\lambda_p = 0.25$	47.02	84.95	84.94	0.717	0.794	0.717	0.680	0.729	80.1	76.0	80.7
$\lambda_p = 0.5$	47.08	85.04	85.03	0.716	0.795	0.722	0.686	0.731	80.6	76.6	80.9
$\lambda_p = 0.75$	47.11	85.09	85.09	0.715	0.796	0.725	0.689	0.733	80.9	77.0	81.1
$\lambda_p = 1.0$ **(default)**†	47.13	85.12	85.12	0.714	0.796	0.727	0.691	0.734	81.1	77.2	81.2
$\lambda_p = 1.5$	47.12	85.11	85.10	0.714	0.796	0.726	0.690	0.734	81.0	77.1	81.1
$\lambda_p = 2.0$	47.11	85.08	85.08	0.715	0.796	0.725	0.689	0.733	80.9	76.9	81.1
w/o consensus term†	47.03	84.92	84.90	0.718	0.793	0.721	0.683	0.704	80.5	76.4	78.3
CMU-MOSEI											
$\lambda_p = 0^\dagger$	55.1	85.45	85.38	0.535	0.768	0.696	0.654	0.707	80.0	75.8	81.1
$\lambda_p = 0.25$	55.2	85.55	85.49	0.533	0.770	0.705	0.664	0.712	80.9	76.9	81.6
$\lambda_p = 0.5$	55.2	85.61	85.56	0.532	0.771	0.710	0.670	0.715	81.5	77.5	82.0
$\lambda_p = 0.75$	55.3	85.65	85.60	0.531	0.772	0.713	0.673	0.718	81.8	78.0	82.2
$\lambda_p = 1.0$ **(default)**†	55.3	85.67	85.61	0.531	0.772	0.715	0.676	0.719	82.0	78.2	82.3
$\lambda_p = 1.5$	55.3	85.66	85.60	0.531	0.772	0.714	0.675	0.719	81.9	78.1	82.2
$\lambda_p = 2.0$	55.3	85.64	85.58	0.531	0.772	0.712	0.673	0.718	81.8	77.9	82.2
w/o consensus term†	55.2	85.50	85.44	0.534	0.769	0.709	0.668	0.685	81.2	77.2	79.0

Table 4: Sensitivity of reliability-aware routing to the log-prior injection strength λ_p and to the presence of the consensus feature in the expert prior. Unless stated otherwise, all other hyperparameters are kept at their defaults, including $\lambda_{\text{PID}} = \lambda_{\text{syn}} = 0.05$. Robustness columns report normalized AUC (nAUC; higher is better) over the Noise/Missing/Conflict severity grids. Rows marked † are directly reported in Table 3.

D Computational Overhead

All experiments use a single NVIDIA A100 (40GB) with batch size 32 for 60 epochs and AdamW. Training uses four views per step ($p \in 0, -T, -A, -V$): masked views replace the corresponding modality encoder output with a MASK vector (no re-encoding), unimodal encoders run once, routing/probe quantities are formed on the full view and reused across views, and the fusion module plus experts run once per view. Inference uses only the full view (no masked views, no DIR), matching a single MoE forward pass.

Training steps and view-level forwards (fusion+experts). CMU-MOSI (train=1,284): 41 steps/epoch, 2,460 total steps; 9,840 view-forwards (baseline single-view: 2,460). CMU-MOSEI (train=16,326): 511 steps/epoch, 30,660 total steps; 122,640 view-forwards (baseline single-view: 30,660). **Compute overhead.** Relative to single-view training, fusion+experts forward/backward compute is $4.0\times$ while unimodal encoders and router/probes remain $1.0\times$; overall training step-time multiplier is $2.80\times$.

Setting	A7↑	A2↑	F1↑	MAE↓	C↑	N(C)↑	M(C)↑	Cf(C)↑	N(A2)↑	M(A2)↑	Cf(A2)↑
CMU-MOSI											
w/o DIR ($\lambda_{\text{PID}}=0,\ \lambda_{\text{syn}}=0$)†	46.33	84.22	84.18	0.734	0.784	0.694	0.652	0.700	77.8	72.9	78.6
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.01$	46.55	84.45	84.42	0.729	0.787	0.700	0.660	0.706	78.7	73.8	79.1
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.03$	46.93	84.84	84.83	0.720	0.792	0.711	0.674	0.717	80.0	75.5	80.0
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.05$ **(default)**†	47.13	85.12	85.12	0.714	0.796	0.727	0.691	0.734	81.1	77.2	81.2
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.08$	47.09	85.08	85.07	0.715	0.795	0.725	0.689	0.732	80.9	76.9	81.1
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.10$	47.05	85.03	85.03	0.716	0.794	0.724	0.687	0.731	80.8	76.8	80.9
$\lambda_{\text{PID}}=0.02,\ \lambda_{\text{syn}}=0.05$	46.88	84.76	84.75	0.721	0.793	0.716	0.680	0.719	80.5	76.5	79.8
$\lambda_{\text{PID}}=0.05,\ \lambda_{\text{syn}}=0.02$	46.83	84.71	84.70	0.722	0.792	0.714	0.677	0.724	80.2	75.9	80.4
$\lambda_{\text{PID}}=0.02,\ \lambda_{\text{syn}}=0.08$	46.90	84.78	84.77	0.721	0.793	0.717	0.681	0.721	80.6	76.5	80.0
$\lambda_{\text{PID}}=0.08,\ \lambda_{\text{syn}}=0.02$	46.87	84.75	84.74	0.722	0.792	0.716	0.679	0.725	80.3	76.0	80.5
w/o L_{PID}†	46.63	84.52	84.50	0.726	0.788	0.715	0.678	0.709	79.9	75.6	78.9
w/o L_{syn}†	46.73	84.62	84.60	0.724	0.790	0.707	0.664	0.717	79.1	74.6	79.8
w/o $L_{\text{syn-off}}$†	46.88	84.78	84.76	0.721	0.791	0.709	0.657	0.724	79.4	74.2	80.2
w/o $L_{\text{syn-res}}$†	46.43	84.32	84.30	0.731	0.786	0.716	0.682	0.725	80.0	76.0	80.4
w/o masked views†	46.23	84.12	84.08	0.737	0.782	0.690	0.646	0.695	77.4	72.2	78.2
CMU-MOSEI											
w/o DIR ($\lambda_{\text{PID}}=0,\ \lambda_{\text{syn}}=0$)†	54.4	84.70	84.62	0.548	0.758	0.684	0.635	0.688	78.4	73.4	79.4
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.01$	54.6	84.92	84.85	0.544	0.761	0.690	0.643	0.694	79.3	74.3	79.9
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.03$	55.0	85.31	85.24	0.537	0.766	0.701	0.658	0.706	80.6	76.1	80.9
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.05$ **(default)**†	55.3	85.67	85.61	0.531	0.772	0.715	0.676	0.719	82.0	78.2	82.3
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.08$	55.2	85.62	85.56	0.532	0.771	0.713	0.674	0.717	81.8	77.9	82.1
$\lambda_{\text{PID}}=\lambda_{\text{syn}}=0.10$	55.2	85.56	85.50	0.533	0.770	0.712	0.672	0.716	81.6	77.7	82.0
$\lambda_{\text{PID}}=0.02,\ \lambda_{\text{syn}}=0.05$	55.0	85.34	85.28	0.536	0.768	0.705	0.664	0.708	81.0	77.3	80.9
$\lambda_{\text{PID}}=0.05,\ \lambda_{\text{syn}}=0.02$	55.0	85.34	85.28	0.536	0.768	0.705	0.663	0.713	81.0	76.7	81.4
$\lambda_{\text{PID}}=0.02,\ \lambda_{\text{syn}}=0.08$	55.0	85.38	85.32	0.535	0.769	0.707	0.666	0.709	81.2	77.2	81.1
$\lambda_{\text{PID}}=0.08,\ \lambda_{\text{syn}}=0.02$	55.0	85.38	85.32	0.535	0.769	0.707	0.665	0.714	81.2	76.8	81.6
w/o L_{PID}†	54.8	85.05	84.98	0.540	0.763	0.704	0.662	0.693	80.9	76.6	79.6
w/o L_{syn}†	54.9	85.10	85.02	0.539	0.765	0.694	0.646	0.706	79.8	74.9	80.7
w/o $L_{\text{syn-off}}$†	55.0	85.38	85.30	0.536	0.767	0.696	0.644	0.710	80.1	74.8	81.0
w/o $L_{\text{syn-res}}$†	54.6	84.92	84.84	0.545	0.762	0.704	0.664	0.710	80.8	76.8	81.4
w/o masked views†	54.3	84.60	84.52	0.550	0.756	0.679	0.631	0.683	77.8	72.6	78.7

Table 5: Sensitivity to DIR weights λ_{PID} and λ_{syn}, including ratio sweeps and component controls. Unless stated otherwise, λ_p is fixed to the default 1.0. Robustness is summarized by nAUC over Noise/Missing/Conflict. Rows marked † are directly reported in Table 3.

Dataset	Params	FP32 MiB	FP16 MiB	Train-state MiB
CMU-MOSI	1,522,299	5.81	2.90	23.23
CMU-MOSEI	12,011,259	45.82	22.91	183.28

Table 6: Parameter count and parameter-related memory footprint (MiB; 1 MiB $= 2^{20}$ bytes). Train-state uses mixed precision with AdamW (16 bytes/parameter: fp16 weights, fp32 master, fp16 grads, fp32 m/v).