RIDER-MoE: Reliability-Conditioned Interaction Routing for Robust Multimodal Sentiment Analysis

Anonymous authors
Paper under double-blind review

Abstract

Multimodal sentiment analysis needs to retain strong performance with complete inputs while remaining robust to noisy, missing, or conflicting modalities. We address this goal within a unified framework through reliability-conditioned interaction routing, adapting the mixture of interaction pathways rather than relying solely on modality reweighting. We present RIDER-MoE, a reliability-aware 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 design is motivated by Partial Information Decomposition (PID). We introduce masked-view Disentangled Interaction Regularization (DIR) to encourage distinct U/R/S behaviors during training while retaining single-pass inference. Lightweight unimodal sentiment probes provide uncertainty and consensus cues that guide expert-level priors, softly biasing a router driven by the fused representation. Across CMU-MOSI, CMU-MOSEI, and CH-SIMS, RIDER-MoE remains competitive with strong baselines on clean full-modality test sets and achieves the highest aggregate normalized AUC for Acc-2 and Corr among the evaluated methods under controlled noise, missing-modality, and cross-modal conflict stress tests. Ablations and routing diagnostics support the complementary roles of reliability-aware routing and masked-view expert regularization.

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., 2024b). 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 & Li, 2025) achieving impressive accuracy on benchmarks such as MOSI and MOSEI (Zadeh et al., 2016; Bagher Zadeh et al., 2018). These models capture cross-modal dependencies to integrate information from text, audio, and vision.

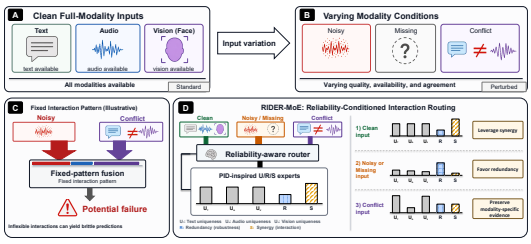

Figure 1: Motivation for reliability-conditioned interaction routing. Top: clean full-modality inputs and variations in modality quality, availability, and agreement, evaluated through controlled stress tests. Bottom: a fixed interaction pattern versus sample-wise routing among PID-inspired U/R/S experts.

However, strong performance with complete multimodal inputs does not by itself establish robustness when modality quality, availability, or agreement changes. Practical concerns such as noisy audio, unavailable visual cues, and inconsistent sentiment evidence motivate studying these variations (Poria et al., 2020). Our goal is to retain competitive full-modality performance while adapting to degraded or conflicting inputs within a unified fusion framework. As illustrated in **Figure 1 (Top)**, we investigate this goal on standard test sets and through controlled noise, missing-modality, and cross-modal conflict stress tests.

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.

Existing studies explore adaptive fusion and multimodal adapters (Wu et al., 2025; Chen et al., 2025). Our focus is not only on *which modalities to trust*, but also on *which interaction pathways to emphasize for each input*. As illustrated in **Figure 1 (Bottom-Left)**, an interaction pattern that does not adapt to modality reliability or cross-modal agreement can become brittle. When a modality becomes unreliable, forcing synergistic fusion can entangle corrupted evidence with clean cues; under cross-modal conflict, the same interaction can obscure informative modality-specific signals. This motivates *sample-wise* interaction routing, together with training signals that encourage different pathways to remain behaviorally distinct.

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 *modality-uniqueness* experts (U_T, U_A, U_V) intended to preserve modality-specific evidence, a *redundancy* expert (R) designed to remain stable under single-modality removal, and a *synergy* expert (S) designed to exploit joint cues that complement the other pathways. This expert factorization is *inspired* by Partial Information Decomposition (PID) (Williams & Beer, 2010; Wollstadt et al., 2023). To give these roles explicit training signals, we introduce masked-view Disentangled Interaction Regularization (DIR; Appendix A), which encourages role-specific invariance for U/R experts and trains S to provide a residual correction on the full view while reducing its output under masking. These constraints shape expert behavior without estimating numerical PID quantities.

A **Reliability-Aware Router** then uses each modality's uncertainty (via entropy) and cross-modal agreement (via a consensus score) to adjust the expert mixture *per sample* through soft expert priors. The design targets three complementary routing behaviors. For **clean and consistent** inputs, the router emphasizes the *Synergy* expert to exploit useful joint cues. For **noisy or missing** modalities, it shifts emphasis away from potentially brittle synergistic interactions and toward the *Redundancy* expert, favoring evidence that is trained to persist under modality removal. For **cross-modal conflicts**, agreement can drop even when unimodal confidence remains high (e.g., positive text paired with negative audio); the router then favors the *Uniqueness* pathways to preserve modality-specific evidence rather than forcing contradictory cues into a synergistic prediction. Together, DIR and reliability-aware routing couple *expert specialization* with *sample-wise strategy selection* within a unified fusion framework.

Our contributions are:

- **Interaction-level robustness formulation.** We distinguish modality degradation from cross-modal conflict and formulate robust MSA as reliability-conditioned routing among modality-unique, redundant, and synergistic interaction pathways, extending reliability modeling from modality weighting to interaction-strategy selection.
- **RIDER-MoE with masked-view role disentanglement.** We introduce a PID-inspired mixture-of-experts model whose U/R/S experts are softly routed using unimodal uncertainty and cross-modal consensus. DIR provides role-specific masked-view constraints to encourage U/R/S specialization while retaining single-pass inference.
- **Controlled robustness evaluation.** On CMU-MOSI, CMU-MOSEI, and CH-SIMS, RIDER-MoE maintains competitive clean full-modality performance and achieves the highest aggregate robustness nAUC for Acc-2 and Corr among the evaluated methods. Ablations and routing diagnostics support the roles of interaction-aware routing and DIR, complemented by a cost–robustness analysis.

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.

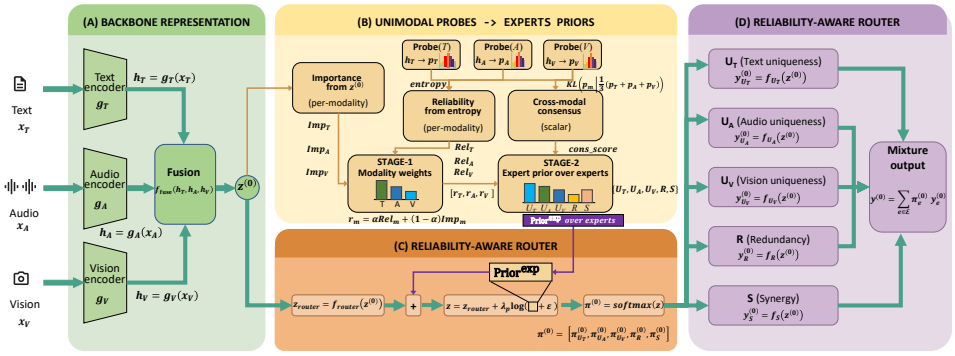

Figure 2: Inference-time architecture of RIDER-MoE with a reliability-aware router over *PID-inspired* U/R/S experts.

Robustness via denoising and reconstruction. Robust MSA has been approached via denoising noisy modalities (Li & 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 & 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 cross-modal consensus to bias routing; DIR further encourages role-specific expert behavior 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 predic-

tion 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 a larger confidence-based weight. 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 relative modality weights r_m for expert-prior construction and uniqueness regularization. Separately, $\bar{r} = \exp(-\frac{1}{3}\sum_{m \in \mathcal{M}} H_m)$ aggregates unnormalized probe confidence for weighting the redundancy and synergy losses; it is not the mean of r_m. 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 $\mathrm{Prior}^{\text{exp}}$ 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 $\lambda_p \log(\mathrm{Prior}_e^{\text{exp}})$ 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 constrains scalar expert outputs

$\{y_e^{(p)}\}$ across masked views, using the relative modality weights r_m and the entropy-based joint confidence proxy $\bar{r}$.

3.4.1 PID-INSPIRED INVARIANCE FOR UNIQUENESS AND REDUNDANCY

The first component, $\mathcal{L}_{\text{PID}}$, 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. The penalty discourages output changes under these specified removals; it does not require the expert to depend only on modality m.

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. Let $\mathcal{P} = \{0, -T, -A, -V\}$. The compact expert-training losses are

$$\mathcal{L}_{\text{uni},m} = \frac{1}{|P_m| - 1} \sum_{p \in P_m \setminus \{0\}} \left(y_{U_m}^{(0)} - y_{U_m}^{(p)}\right)^2, \quad \mathcal{L}_{\text{uni}} = \sum_{m \in \mathcal{M}} r_m \mathcal{L}_{\text{uni},m}. \tag{3}$$

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

$$\mathcal{L}_{\text{PID}} = \mathcal{L}_{\text{uni}} + \mathcal{L}_{\text{red}}. \tag{5}$$

Appendix B.3 gives the expanded definitions and interpretation.

3.4.2 SYNERGY REGULARIZATION

The second component, $\mathcal{L}_{\text{syn}}$, gives S a full-view correction role: its output is penalized on the three single-modality masked views, while its routed full-view contribution is trained to fit the current non-synergy residual.

Concretely, with $\mathcal{P}^- = \{-T, -A, -V\}$, the off-under-masking and residual-fit terms are

$$\mathcal{L}_{\text{syn-off}} = \bar{r}^{\,3} \cdot \frac{1}{3} \sum_{p \in \mathcal{P}^-} \left(y_S^{(p)}\right)^2. \tag{6}$$

$$\hat{y}_{\setminus S}^{(0)} = \sum_{e \in \mathcal{E} \setminus \{S\}} \pi_e^{(0)} y_e^{(0)}, \quad \delta = y - \text{sg}\left(\hat{y}_{\setminus S}^{(0)}\right), \quad \mathcal{L}_{\text{syn-res}} = \bar{r}^{\,3} \left(\pi_S^{(0)} y_S^{(0)} - \delta\right)^2. \tag{7}$$

$$\mathcal{L}_{\text{syn}} = \mathcal{L}_{\text{syn-off}} + \mathcal{L}_{\text{syn-res}}. \tag{8}$$

Here $\text{sg}(\cdot)$ blocks gradients through the target branch within each update. The factor $\bar{r}^3 = \exp(-\sum_{m \in \mathcal{M}} H_m)$ gives greater weight to examples with lower aggregate probe entropy, not necessarily more accurate or more mutually consistent predictions. The weights remain differentiable; Appendix A explains the resulting gradient paths, and Appendix B.3 expands the losses.

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:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{cls}} \sum_{m \in \mathcal{M}} \mathcal{L}_{\text{cls}}^{(m)} + \lambda_{\text{PID}} \mathcal{L}_{\text{PID}} + \lambda_{\text{syn}} \mathcal{L}_{\text{syn}}. \tag{9}$$

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}).

Table 1: Clean full-modality performance on CMU-MOSI and CMU-MOSEI.

Methods	MOSI					MOSEI				
	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑
TFN (Zadeh et al., 2017)	34.4	79.0	79.1	0.947	0.673	51.6	81.8	81.7	0.573	0.714
LMF (Liu et al., 2018)	33.8	79.1	79.1	0.950	0.651	51.5	83.4	83.3	0.576	0.717
MulT (Tsai et al., 2019b)	40.4	83.4	83.5	0.846	0.725	52.6	83.5	83.6	0.564	0.731
MMIM (Han et al., 2021)	45.8	84.6	84.5	0.717	0.786	50.1	83.6	83.5	0.580	0.756
MISA (Hazarika et al., 2020)	41.3	83.5	83.5	0.776	0.778	52.0	84.6	84.6	0.557	0.751
CCIN-SA (Yang et al., 2025)	43.3	83.9	83.9	0.810	0.731	54.1	84.7	84.9	0.545	0.736
Self-MM (Yu et al., 2021)	46.4	84.6	84.6	0.717	0.793	53.6	85.0	85.0	0.533	0.766
TMBL (Huang et al., 2024)	36.3	83.8	84.2	0.867	0.762	52.4	85.8	85.9	0.545	0.766
EMOE (Fang et al., 2025)	45.8	**85.2**	**85.2**	0.725	0.792	53.1	85.1	84.9	0.543	0.759
DLF (Wang et al., 2025)	**47.1**	85.1	85.0	0.731	0.781	53.9	85.4	85.3	0.536	0.764
RIDER-MoE (**Ours**)	46.9	84.8	85.1	**0.709**	**0.796**	**54.3**	**85.9**	**86.1**	**0.531**	**0.772**

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 inference is a single full-view MoE forward pass with one router evaluation and five expert heads. We evaluate the efficiency and cost–robustness trade-off in Section 4.3. Detailed pseudo-code for training and inference is given in Algorithm 1 in Appendix B.4.

4 Experiments

Evaluation protocol. We evaluate RIDER-MoE on CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020). The experiments address three questions: **whether RIDER-MoE preserves clean full-modality performance**, **whether it improves robustness under test-time degradation and cross-modal disagreement**, and **whether the gains come from reliability-aware U/R/S routing rather than extra capacity**. For CMU-MOSI and CMU-MOSEI, we report Acc-7, Acc-2, F1, MAE, and Corr. For CH-SIMS, we report Acc-2, Acc-5, F1, MAE, and Corr. Robustness results use Acc-2 and Corr, summarized by normalized AUC (nAUC). Implementation details, stress-test protocols, and diagnostic metrics are provided in Appendix C.

4.1 Clean Full-Modality Performance

Tables 1 and 2 first establish a clean full-modality reference. On CMU-MOSI, RIDER-MoE achieves the best MAE and Corr, while staying close to the strongest classification baselines. On CMU-MOSEI, it ranks first on all reported metrics. On CH-SIMS, it also achieves the best Acc-2, Acc-5, F1, MAE, and Corr. These results show that the reliability-aware design preserves normal-case accuracy. The robustness experiments below therefore test stability under stress, not recovery from weak clean performance.

Table 2: Clean full-modality performance on CH-SIMS.

Methods	Acc-2↑	Acc-5↑	F1↑	MAE↓	Corr↑
TFN (Zadeh et al., 2017)	78.3	39.3	78.6	0.432	0.591
LMF (Liu et al., 2018)	77.7	40.5	77.8	0.441	0.575
MulT (Tsai et al., 2019b)	78.2	40.0	78.5	0.442	0.581
Self-MM (Yu et al., 2021)	78.6	43.1	78.6	0.411	0.601
TMBL (Huang et al., 2024)	79.1	41.5	78.7	0.429	0.592
EMOE (Fang et al., 2025)	75.7	40.0	75.9	0.449	0.519
DLF (Wang et al., 2025)	80.1	39.8	78.8	0.446	0.563
RIDER-MoE (**Ours**)	**81.2**	**44.3**	**81.1**	**0.405**	**0.615**

4.2 Robustness under Test-Time Stressors

Protocol and baselines. We stress-test models with Noise (N), Missingness (M), and Cross-modal Conflict (Cf). Noise corrupts modality content; Missingness replaces selected modality representations with mask vectors; Conflict pairs the original text and label with mismatched non-text modalities. Stressors are test-time only. All methods see the same corrupted examples at

Table 3: **Robustness nAUC under controlled test-time stressors.** Each cell reports Acc-2 nAUC / Corr nAUC; higher is better. N, M, and Cf denote Noise, Missingness, and Cross-modal Conflict. Bold marks the best value for each metric within each dataset–stressor column. Avg macro-averages the nine dataset–stressor cells.

Method	MOSI			MOSEI			CH-SIMS			Avg
	N	M	Cf	N	M	Cf	N	M	Cf	
Full-modality MSA baselines										
Self-MM (Yu et al., 2021)	79.2/0.705	74.6/0.645	78.4/0.690	79.9/0.672	75.1/0.618	78.7/0.660	72.4/0.535	67.8/0.500	70.5/0.520	75.2/0.616
TMBL (Huang et al., 2024)	78.0/0.680	73.8/0.628	77.2/0.660	80.7/0.674	76.0/0.620	79.0/0.662	72.9/0.525	68.1/0.492	70.8/0.508	75.2/0.605
EMOE (Fang et al., 2025)	80.0/0.708	75.3/0.653	79.7/0.710	80.3/0.682	75.8/0.632	80.0/0.690	69.5/0.455	65.5/0.420	68.4/0.440	74.9/0.599
DLF (Wang et al., 2025)	79.7/0.700	75.1/0.648	80.2/0.718	81.0/0.688	76.1/0.638	80.8/0.704	74.1/0.510	69.2/0.480	73.2/0.512	76.6/0.622
Robustness-oriented baselines										
GCNet (Lian et al., 2023)	78.6/0.680	76.0/0.655	76.8/0.665	79.8/0.668	76.9/0.648	78.2/0.655	71.0/0.500	68.5/0.485	68.9/0.492	75.0/0.605
CorrKD (Li et al., 2024a)	79.0/0.690	76.6/0.665	77.5/0.670	80.5/0.678	77.5/0.657	78.9/0.665	71.8/0.512	69.0/0.490	69.3/0.500	75.6/0.614
LNLN (Zhang et al., 2024)	80.1/0.710	77.0/0.675	78.8/0.700	81.3/0.696	78.2/0.668	80.1/0.690	75.0/0.552	71.0/0.526	73.6/0.542	77.2/0.640
TMDC (Zhuang et al., 2026)	**80.8**/0.716	**77.4/0.686**	78.3/0.688	**81.7**/0.704	**78.5/0.674**	79.7/0.682	74.5/0.545	70.7/0.520	72.1/0.525	77.1/0.638
Proposed method										
RIDER-MoE	80.7/**0.722**	77.3/0.684	**81.0/0.733**	81.6/**0.706**	78.4/0.672	**82.0/0.718**	**75.2/0.565**	**72.0/0.540**	**75.0/0.560**	**78.1/0.656**

each severity, and none is tuned on stressed test inputs. We sweep $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, $m \in \{0, 0.2, 0.4, 0.6, 0.8\}$, and $c \in \{0, 0.25, 0.5, 0.75, 1.0\}$, and summarize each curve by nAUC (Appendix C.3). Baselines include strong full-modality MSA models and robustness-oriented methods for incomplete, noisy, or degraded inputs. Methods follow official recipes when available and share the same feature interface and validation protocol. Since several robustness baselines are not designed for semantic mismatch, Conflict results are treated as diagnostic transfer.

Stress-test results. Table 3 gives the main robustness result. RIDER-MoE achieves the strongest aggregate score, with 78.1 Acc-2 nAUC and 0.656 Corr nAUC, improving over the best non-RIDER average by 0.9 and 0.016. The comparison remains competitive: TMDC slightly leads on several degradation-only cells, especially Noise and Missingness on CMU-MOSI and CMU-MOSEI. The clearest separation appears under Cross-modal Conflict, where each modality can remain plausible while their joint semantics disagree. RIDER-MoE obtains the best Conflict nAUC on all three datasets and both metrics, improving over the strongest non-RIDER baseline by 0.8/0.015 on CMU-MOSI, 1.2/0.014 on CMU-MOSEI, and 1.4/0.018 on CH-SIMS. Figure 3 shows the same trend: RIDER-MoE decays more slowly as conflict severity increases. Overall, it ranks first in 12 of 18 dataset–stressor–metric comparisons.

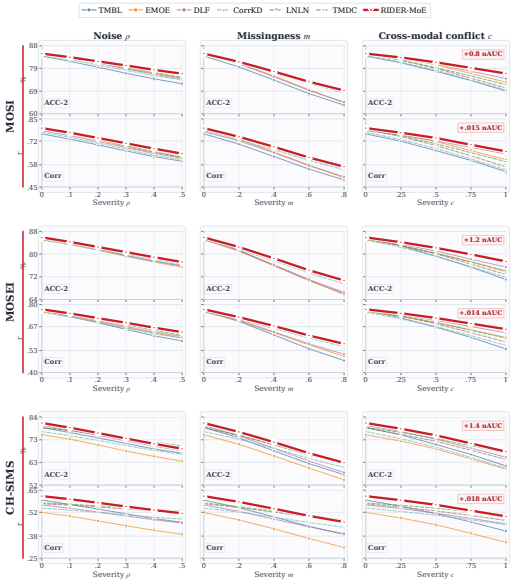

Figure 3: **Performance–severity curves under test-time stressors.** Acc-2 and Corr are reported on CMU-MOSI, CMU-MOSEI, and CH-SIMS under Noise, Missingness, and Cross-modal Conflict.

These results match the intended routing behavior. Noise and Missingness mainly reduce reliability, so redundancy-oriented fallback is effective. Conflict instead reduces agreement while unimodal confidence may remain high. RIDER-MoE uses uncertainty and agreement to suppress risky Synergy, favor Redundancy under degraded evidence, and shift toward Uniqueness under disagreement. The largest gains therefore appear when robustness requires interaction-type selection, not only denoising or completion.

4.3 Cost–Robustness Trade-off

Cost controls. We ask whether the robustness gain requires full DIR, and how much cost it adds. The controls isolate capacity, routing, generic robustness augmentation, DIR training cost, and

Table 4: **Cost–robustness trade-off.** Metric cells report Acc-2/Corr. Clean Avg macro-averages CMU-MOSI, CMU-MOSEI, and CH-SIMS. Robust Avg nAUC averages Noise, Missingness, and Cross-modal Conflict. Train cost is normalized to one-view RIDER-MoE without DIR; inference latency is normalized to the capacity-matched single-head model.

Variant	Train Cost	Infer Latency	Clean Avg Acc-2/Corr	Robust Avg nAUC			
				MOSI	MOSEI	CH-SIMS	Overall
Capacity and routing controls							
Capacity-matched single-head	0.86×	1.00×	82.1/0.701	74.9/0.656	75.2/0.648	69.3/0.498	73.1/0.601
RIDER-MoE w/o DIR	1.00×	1.18×	83.0/0.714	76.5/0.675	77.4/0.665	71.2/0.522	75.0/0.621
Simple robustness controls							
w/o DIR + Modality Dropout	1.01×	1.18×	82.7/0.711	77.2/0.685	78.1/0.676	71.9/0.533	75.7/0.631
w/o DIR + Noise Augmentation	1.02×	1.18×	82.7/0.711	77.0/0.680	78.0/0.672	71.7/0.528	75.6/0.627
DIR cost controls							
RIDER-MoE + Curriculum DIR	1.30×	1.18×	83.4/0.720	78.1/0.699	79.3/0.689	72.8/0.541	76.7/0.643
RIDER-MoE + Stochastic DIR	1.60×	1.18×	83.7/0.724	78.8/0.706	80.0/0.695	73.5/0.550	77.4/0.650
RIDER-MoE + Full DIR	2.80×	1.18×	**84.0/0.728**	**79.7/0.713**	**80.7/0.699**	**74.1/0.555**	**78.1/0.656**
Full DIR + Top-2 inference	2.80×	1.07×	83.6/0.723	79.2/0.708	80.2/0.694	73.4/0.548	77.6/0.650

conditional inference. We compare a capacity-matched single-head model, RIDER-MoE without DIR, no-DIR variants with modality dropout or noise augmentation, and three DIR schedules. Curriculum DIR applies full four-view DIR only in the last 10 of 60 epochs. Stochastic DIR uses the full view plus one randomly sampled masked view per mini-batch. Full DIR uses all three single-modality masked views throughout training. Top-2 inference uses the Full-DIR checkpoint but executes only the two highest-weighted experts at test time.

Cost results. Table 4 separates the sources of improvement. Routing is not just extra head capacity: RIDER-MoE without DIR improves over the capacity-matched single-head control on both clean and robust averages. Generic augmentation helps but is insufficient: modality dropout and noise augmentation improve the no-DIR model, yet both remain below Full DIR. The remaining gain comes from masked-view role shaping, which constrains how U/R/S experts behave when views are removed. The cost is tunable: Curriculum DIR lowers training cost from 2.80× to 1.30×, Stochastic DIR recovers most of the Full-DIR robustness at 1.60× cost, and Top-2 inference reduces latency from 1.18× to 1.07× with only a small robustness drop.

4.4 Ablation and Mechanistic Analysis

Ablation setup. We test why the robustness gain occurs. The ablations ask whether performance comes from interaction-type routing and DIR constraints, rather than added heads, prior injection alone, or generic regularization. All variants share the same encoders, fusion backbone, optimizer, validation protocol, and stress-test grids. Stressors are applied only at test time. Table 5 reports clean Acc-2/Corr and robust nAUC under Noise (N), Missingness (M), and Cross-modal Conflict (Cf), macro-averaged over CMU-MOSI, CMU-MOSEI, and CH-SIMS.

Table 5: **Component ablation.** Each cell reports Acc-2/Corr. Clean is the uncorrupted-test macro average. N, M, and Cf are robust nAUC under Noise, Missingness, and Cross-modal Conflict. Robust Avg macro-averages all dataset–stressor cells.

Variant	Clean	Robust nAUC			
		N	M	Cf	Avg
RIDER-MoE	**84.0/0.728**	**79.2/0.664**	**75.9/0.632**	**79.3/0.670**	**78.1/0.656**
w/o prior (λ_p=0)	83.6/0.723	77.2/0.641	73.9/0.608	78.3/0.657	76.5/0.635
w/o consensus	83.8/0.725	78.7/0.658	75.8/0.629	76.8/0.641	77.1/0.642
w/o DIR	83.0/0.714	76.2/0.632	71.5/0.586	77.5/0.644	75.0/0.621
w/o $\mathcal{L}_{\text{PID}}$	83.4/0.720	78.4/0.654	74.5/0.619	77.6/0.651	76.8/0.642
w/o $\mathcal{L}_{\text{syn}}$	83.4/0.720	77.6/0.645	73.7/0.608	78.1/0.655	76.4/0.636
single-head	82.1/0.701	74.1/0.611	69.4/0.561	75.9/0.630	73.1/0.601

Component attribution. Table 5 supports three conclusions. First, robustness is not a capacity artifact. Replacing the U/R/S experts with a capacity-matched single head drops Robust Avg from 78.1/0.656 to 73.1/0.601. Second, routing alone is insufficient. Removing DIR while keeping the expert structure and router causes a large drop, especially under Missingness. This shows that masked-view role shaping is needed to keep

experts behaviorally distinct under degraded inputs. Third, the priors are stressor-specific. Removing prior injection lowers overall robustness, while removing only consensus mainly hurts Conflict. This matches the design: Noise and Missingness reduce reliability, whereas Conflict reduces agreement even when unimodal predictions remain confident.

The two DIR losses play complementary roles. Removing $\mathcal{L}_{\text{PID}}$ weakens Conflict robustness, consistent with degraded uniqueness and redundancy invariance. Removing $\mathcal{L}_{\text{syn}}$ hurts Noise and Missingness more, consistent with weaker suppression of Synergy under masked or degraded views. Together with the no-DIR and single-head variants, these results show that robustness requires both interaction-specialized experts and DIR constraints that preserve their roles. Appendix C.6 reports sensitivity around the default λ_p and λ_D values.

Mechanistic diagnostics. Figure 4 tracks average full-view routing weights as test-time stress increases.

On clean inputs, the router assigns substantial mass to Synergy. Under Noise and Missingness, π_S decreases while π_R increases, especially under Missingness, consistent with fallback to the route encouraged to persist under modality removal. Conflict produces a different pattern: π_U increases, π_S drops, and π_R remains relatively stable. These responses are consistent with the intended distinction between modality degradation and disagreement, although mixture weights alone do not identify the information used by each expert.

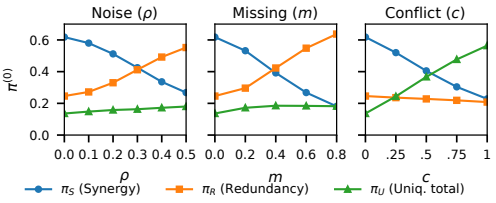

Figure 4: **Router adaptation under stress.** Average full-view routing weights under Noise, Missingness, and Cross-modal Conflict as severity increases. Uniqueness is aggregated as $\pi_U = \pi_{U_T} + \pi_{U_A} + \pi_{U_V}$.

Figure 5 checks masked-view expert behavior without changing normal full-view inference. Lower ΔU_{avg} and $\text{Var}(R)$ indicate greater uniqueness and redundancy stability under the specified masks; lower $\text{Off}(S)$ indicates stronger suppression of masked-view synergy outputs. Higher $\text{corr}(\pi_S^{(0)} y_S^{(0)}, \delta)$ indicates stronger alignment with the current non-synergy residual, which does not isolate interaction-only information.

The full model shows low uniqueness inconsistency, low redundancy variance, low synergy leakage, and high residual alignment. Removing $\mathcal{L}_{\text{PID}}$ mainly degrades uniqueness and redundancy diagnostics, while removing $\mathcal{L}_{\text{syn}}$ mainly degrades synergy diagnostics; removing DIR degrades multiple checks at once. These behavioral checks support role shaping without identifying PID atoms; definitions and interpretive limits are given in Appendices C.4 and A.

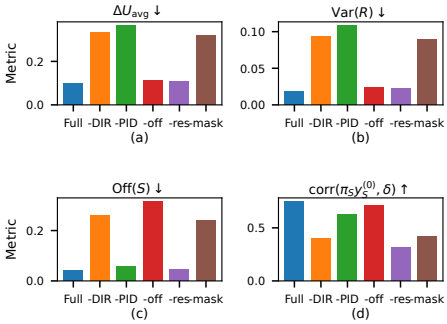

Figure 5: **Behavioral checks for DIR.** Lower is better for uniqueness inconsistency ΔU_{avg}, redundancy variance $\text{Var}(R)$, and synergy leakage $\text{Off}(S)$; higher is better for residual alignment $\text{corr}(\pi_S^{(0)} y_S^{(0)}, \delta)$.

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 CMU-MOSI, CMU-MOSEI, and CH-SIMS show that RIDER-MoE maintains competitive clean performance while improving robustness under noise, missing-modality, and conflict stress tests. These findings support reliability-conditioned interaction routing. Future work includes evaluating under more realistic distribution shifts and extending reliability-aware routing to additional modalities and tasks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript drafting and polishing, figure refinement, and reproducing baseline implementations from prior work. These tools were not used to originate the core research ideas, design the experiments, or generate experimental data. All substantive research decisions and scientific conclusions were made by the authors. The authors reviewed the AI-assisted content and code and take full responsibility for the final manuscript and results.

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. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016. doi: 10.1109/WACV.2016.7477553.

K. Chen, L. Xu, and H. Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*, 12(9):1229–1252, 1999. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(99)00043-X. URL https://www.sciencedirect.com/science/article/pii/S089360809900043X.

Kezhou Chen, Shuo Wang, Huixia Ben, Shengeng Tang, and Yanbin Hao. Mixture of multimodal adapters for sentiment analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.),

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), pp. 1822–1833, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.90. URL https://aclanthology.org/2025.naacl-long.90/.

Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17300–17311, 2023. doi: 10.1109/ICCV51070.2023.01591.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep — a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, 2014. doi: 10.1109/ICASSP.2014.6853739.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14314–14324, 2025. doi: 10.1109/CVPR52734.2025.01335.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.723. URL https://aclanthology.org/2021.emnlp-main.723/.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis, 2020. URL https://arxiv.org/abs/2005.03545.

Xilin He, Haijian Liang, Boyi Peng, Weicheng Xie, Muhammad Haris Khan, Siyang Song, and Zitong Yu. Msamba: Exploring multimodal sentiment analysis with state space models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1309–1317, Apr. 2025. doi: 10.1609/aaai.v39i2.32120. URL https://ojs.aaai.org/index.php/AAAI/article/view/32120.

Jiehui Huang, Jun Zhou, Zhenchao Tang, Jiaying Lin, and Calvin Yu-Chian Chen. TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowledge-Based Systems*, 285:111346, 2024. doi: 10.1016/j.knosys.2023.111346. URL https://doi.org/10.1016/j.knosys.2023.111346.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12458–12468, June 2024a.

Qian Li, Lixin Su, Jiashu Zhao, Long Xia, Hengyi Cai, Suqi Cheng, Hengzhu Tang, Junfeng Wang, and Dawei Yin. Text-video retrieval via variational multi-modal hypergraph networks, 2024b. URL https://arxiv.org/abs/2401.03177.

Zuocheng Li and Lishuang Li. t-HNE: A text-guided hierarchical noise eliminator for multimodal sentiment analysis. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2834–2844, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.192/.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8419–8432, 2023. doi: 10.1109/TPAMI.2023.3234553.

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. Quantifying & modeling multimodal interactions: An information decomposition framework. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27351–27393. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/575286a73f238b6516ce0467d67eadb2-Paper-Conference.pdf.

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 Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2256, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1209. URL https://aclanthology.org/P18-1209/.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pp. 169–176, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306416. doi: 10.1145/2070481.2070509. URL https://doi.org/10.1145/2070481.2070509.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research, 2020. URL https://arxiv.org/abs/2005.00357.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*, 14(1):108–132, 2023. doi: 10.1109/TAFFC.2020.3038167.

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.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019a. Association for Computational Linguistics. 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, 2019b. URL https://arxiv.org/abs/1906.00295.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. DLF: Disentangled-language-focused multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21180–21188, Apr. 2025. doi: 10.1609/aaai.v39i20.35416. URL https://ojs.aaai.org/index.php/AAAI/article/view/35416.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information, 2010. URL https://arxiv.org/abs/1004.2515.

Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

Han Wu, Yanming Sun, Yunhe Yang, and Derek F. Wong. Beyond simple fusion: Adaptive gated fusion for robust multimodal sentiment analysis, 2025. URL https://arxiv.org/abs/2510.01677.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L. Ballard, Tianlong Chen, and Qi Long. I2moe: Interpretable multimodal interaction-aware mixture-of-experts, 2025. URL https://arxiv.org/abs/2505.19190.

Li Yang, Junhong Zhong, Teng Wen, and Yuan Liao. CCIN-SA: Composite cross modal interaction network with attention enhancement for multimodal sentiment analysis. *Information Fusion*, 123: 103230, 2025. doi: 10.1016/j.inffus.2025.103230. URL https://doi.org/10.1016/j.inffus.2025.103230.

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 Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.343. URL https://aclanthology.org/2020.acl-main.343/.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, 2021. URL https://arxiv.org/abs/2102.04830.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016. URL https://arxiv.org/abs/1606.06259.

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.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115/.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 55943–55974. Curran Associates, Inc., 2024. doi: 10.52202/079017-1779. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/65a39213d7d0e1eb5d192aa77e77eeb7-Paper-Conference.pdf.

Xueliang Zhao, Mingyang Wang, Yingchun Tan, and Xianjie Wang. Tgmoe: A text guided mixture-of-experts model for multimodal sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 15(8), 2024. doi: 10.14569/IJACSA.2024.01508119. URL http://dx.doi.org/10.14569/IJACSA.2024.01508119.

Miao Zhou, Lina Yang, Thomas Wu, Dongnan Yang, and Xinru Zhang. Dual-path dynamic fusion with learnable query for multimodal sentiment analysis. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11366–11376, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.571. URL https://aclanthology.org/2025.emnlp-main.571/.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22123–22138, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1075. URL https://aclanthology.org/2025.acl-long.1075/.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. TMDC: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and

noisy modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2281–2289, 2026. doi: 10.1609/aaai.v40i3.37212. URL https://doi.org/10.1609/aaai.v40i3.37212.

A Behavioral Interpretation of the PID-Inspired U/R/S Regularization

A.1 PID motivation and scope

Partial Information Decomposition (PID) distinguishes unique, redundant, and synergistic contributions to the information that a set of sources provides about a target (Williams & Beer, 2010; Wollstadt et al., 2023). A numerical PID requires a specified redundancy measure and an associated decomposition. RIDER-MoE uses these concepts to motivate expert roles, without estimating mutual information or identifying PID atoms. Its five scalar experts constitute a predictive architecture, not a complete tri-source information decomposition.

DIR connects these roles to explicit training objectives: U_m is encouraged to remain stable when another modality is removed, R is encouraged to persist across single-modality removals, and S is trained to supply a full-view residual correction with reduced outputs on masked views. This section explains the responses directly encouraged by the losses and their role in reliability-conditioned interaction routing.

A.2 Shared-input experts, masked views, and confidence weights

We follow Sections 3.1–3.3. Let $\mathcal{M} = \{T, A, V\}$, $\mathcal{E} = \{U_T, U_A, U_V, R, S\}$, $\mathcal{P} = \{0, -T, -A, -V\}$, and $\mathcal{P}^- = \{-T, -A, -V\}$. For DIR training, $p = 0$ denotes the complete view and $p = -m$ replaces h_m with MASK_m, leaving the other modality representations unchanged. The fusion module and expert heads are shared across views:

$$\begin{aligned} z^{(p)} &= f_{\text{fuse}}\left(h_T^{(p)}, h_A^{(p)}, h_V^{(p)}\right), \\ y_e^{(p)} &= f_e\left(z^{(p)}\right), \qquad e \in \mathcal{E}. \end{aligned} \tag{10}$$

All experts receive the fused representation; in particular, U_m is not structurally restricted to h_m. The expert names therefore denote intended training roles rather than isolated input connections. The masking interpretation assumes that replacement values are chosen independently of the sample's label and removed content. It concerns removal of a direct input representation, not removal of correlated evidence from the remaining modalities or a causal intervention on the data-generating process. For stochastic replacements, the statements below apply to the realized comparisons.

The full-view weights $\pi^{(0)}$ are reused within the auxiliary DIR losses, with prediction $\hat{y}^{(0)} = \sum_{e\in\mathcal{E}} \pi_e^{(0)} y_e^{(0)}$. Appendix B.1 defines relative modality weights r_m and a separate entropy-based joint confidence proxy:

$$\begin{aligned} \bar{r} &= \exp\left(-\frac{1}{3}\sum_{m\in\mathcal{M}} H_m\right) = \left(\prod_{m\in\mathcal{M}} \exp(-H_m)\right)^{1/3}, \\ \bar{r}^{\,3} &= \exp\left(-\sum_{m\in\mathcal{M}} H_m\right). \end{aligned} \tag{11}$$

Thus $\bar{r}$ is the geometric mean of the unnormalized confidence scores $\exp(-H_m)$, not the arithmetic mean of r_m. The relative weights and consensus score enter expert-prior construction, whereas $\bar{r}$ scales the redundancy and synergy regularizers. A larger $\bar{r}$ reflects lower aggregate probe entropy, not a calibrated probability that all modalities are correct. Confident probes may disagree, so confidence weighting and cross-modal consensus serve distinct purposes.

A.3 BEHAVIORAL ROLES ENCOURAGED BY DIR

A.3.1 UNIQUENESS: STABILITY TO REMOVAL OF OTHER MODALITIES

Let $P_m = \mathcal{P} \setminus \{-m\}$. The uniqueness loss is

$$\begin{aligned} \mathcal{L}_{\text{uni},m} &= \frac{1}{|P_m| - 1} \sum_{p \in P_m \setminus \{0\}} \left(y_{U_m}^{(0)} - y_{U_m}^{(p)}\right)^2, \\ \mathcal{L}_{\text{uni}} &= \sum_{m \in \mathcal{M}} r_m \mathcal{L}_{\text{uni},m}. \end{aligned} \tag{12}$$

This loss penalizes changes in U_m when one of the other modalities is masked, while keeping modality m available. For a particular comparison, zero $\mathcal{L}_{\text{uni},m}$ implies equality of the compared outputs because every summand is nonnegative.

The resulting property is stability under the specified masks, rather than dependence on h_m alone. For example, if audio and vision encode the same varying signal while text is fixed, a predictor can recover that signal from whichever of audio or vision remains available. Such a predictor can satisfy both single-removal equalities for U_T without using text. Accordingly, the loss supplies a modality-indexed invariance bias but does not establish structural modality isolation or identify PID-unique information.

A.3.2 REDUNDANCY: PERSISTENCE ACROSS SINGLE-MODALITY REMOVALS

The redundancy loss is

$$\begin{aligned} \bar{y}_R &= \frac{1}{4} \sum_{p \in \mathcal{P}} y_R^{(p)}, \\ \mathcal{L}_{\text{red}} &= \bar{r}\, \frac{1}{4} \sum_{p \in \mathcal{P}} \left(y_R^{(p)} - \bar{y}_R\right)^2. \end{aligned} \tag{13}$$

It encourages a prediction that persists across the complete view and the three single-modality removals. Expanding the squares gives the equivalent pairwise form

$$\mathcal{L}_{\text{red}} = \frac{\bar{r}}{32} \sum_{p \in \mathcal{P}} \sum_{q \in \mathcal{P}} \left(y_R^{(p)} - y_R^{(q)}\right)^2. \tag{14}$$

For finite probe entropies, $\bar{r} > 0$, so zero loss implies equality of all four outputs.

This is a constraint on cross-view prediction variability. It does not selectively eliminate modality-specific latent factors while preserving shared factors: any direction that changes the output across the compared views can incur a penalty. A constant output also achieves zero variance. Predictive usefulness must therefore be assessed together with the supervised objective and empirical results, rather than inferred from invariance alone.

A.3.3 SYNERGY: SUPPRESSION ON DESIGNATED MASKED VIEWS

The off-under-masking term is

$$\mathcal{L}_{\text{syn}-\text{off}} = \bar{r}^{\,3} \frac{1}{3} \sum_{p \in \mathcal{P}^-} \left(y_S^{(p)}\right)^2. \tag{15}$$

For $\bar{r} > 0$, zero loss implies $y_S^{(p)} = 0$ on each of the three specified masked views. The objective assigns S a full-trimodal-view role rather than a fallback role under these removals. It constrains expert outputs, not the routing probability π_S.

This role is an architectural choice. Synergy can also occur between two modalities, and the loss does not identify all lower-order or higher-order PID components. Moreover, recognizing a masking vector and suppressing an output is not sufficient to establish that the full-view output contains only interaction information. The residual objective below provides the complementary predictive training signal for S.

A.3.4 RESIDUAL FITTING AND GRADIENT ALLOCATION

On the complete view, define

$$\begin{aligned} \hat{y}^{(0)}_{\setminus S} &= \sum_{e \in \mathcal{E} \setminus \{S\}} \pi^{(0)}_e y^{(0)}_e, \\ \delta &= y - \mathrm{sg}\Big(\hat{y}^{(0)}_{\setminus S}\Big), \\ \mathcal{L}_{\mathrm{syn-res}} &= \bar{r}^{\,3}\big(\pi^{(0)}_S y^{(0)}_S - \delta\big)^2. \end{aligned} \tag{16}$$

Here sg preserves forward values and blocks gradients through its argument. The target is the residual of the *current* non-synergy contribution, using the original routing weights without renormalization. It is not the error of an independently optimized model without S, nor the information that the other experts are fundamentally unable to represent.

Let $g = \pi^{(0)}_S y^{(0)}_S$ and $w = \bar{r}^{\,3}$. Since stop-gradient preserves values,

$$g - \delta = \hat{y}^{(0)} - y \qquad \text{in the forward evaluation.} \tag{17}$$

Thus the numerical loss is a confidence-weighted squared prediction error, while its optimization effect depends on the detached target branch. For the computational graph specified above, the back-propagated gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{syn-res}} = 2w(g-\delta)\nabla_\theta g + (g-\delta)^2 \nabla_\theta w. \tag{18}$$

The first term updates the active routed contribution, including the synergy head, the coupled softmax router, and shared upstream modules. The second term updates the entropy-based confidence weight through the probes and their upstream modules. No additional stop-gradient on r_m or $\bar{r}$ is specified in the objective. Stopping the target branch therefore does not freeze shared parameters, the remaining routing probabilities, or subsequent non-synergy predictions.

The confidence weights are learned quantities, so reducing a weighted loss can involve changing either its discrepancy or its weight. The task and probe cross-entropy losses are optimized jointly with these terms; the weighting mechanism does not itself establish confidence calibration. Combining residual fitting with masked-view suppression trains S toward a full-view correction role, without claiming that the residual consists exclusively of synergistic information or that joint optimization reaches a conditional-expectation solution.

A.4 INTERPRETATION OF THE DIAGNOSTICS AND SCOPE

The total DIR objective retains $\mathcal{L}_{\mathrm{PID}} = \mathcal{L}_{\mathrm{uni}} + \mathcal{L}_{\mathrm{red}}$ and $\mathcal{L}_{\mathrm{syn}} = \mathcal{L}_{\mathrm{syn-off}} + \mathcal{L}_{\mathrm{syn-res}}$, combined with the task and probe losses in Eq. (9). The zero-loss implications above describe individual comparisons; they are not claims that training reaches these limits.

The diagnostics in Section 4.4 and Appendix C.4 assess the corresponding behaviors: ΔU_{avg} measures stability under removal of other modalities, $\mathrm{Var}(R)$ measures cross-view variability, $\mathrm{Off}(S)$ measures masked-view output magnitude, and residual correlation measures alignment with the current non-synergy residual. These unweighted checks complement the confidence-weighted training losses. They neither identify PID atoms nor independently establish that every expert is informative; in particular, residual correlation does not distinguish single-modality prediction error from genuinely joint information. Routing probabilities likewise describe mixture allocation, while predictive contributions also depend on expert outputs.

The supported interpretation is that PID motivates complementary expert roles and DIR supplies role-specific prediction-space constraints for sample-wise interaction routing. Their utility is assessed empirically through prediction performance, ablations, and behavioral diagnostics. Robustness under different corruption mechanisms, simultaneous modality removals, or alternative mask representations does not follow from the single-removal constraints alone.

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:

$$\ell_m^{\text{cls}} = W_m^{\text{cls}} h_m + b_m^{\text{cls}} \in \mathbb{R}^K, \tag{19}$$

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

$$p_m(k) = \frac{\exp(\ell_{m,k}^{\text{cls}})}{\sum_{k'=1}^{K} \exp(\ell_{m,k'}^{\text{cls}})}, \quad k = 1, \dots, K. \tag{20}$$

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

$$\mathcal{L}_{\text{cls}}^{(m)} = \text{CE}(p_m, \tilde{y}). \tag{21}$$

B.1.2 Entropy-based modality reliability

From each unimodal distribution p_m, we compute the entropy

$$H_m = -\sum_{k=1}^{K} p_m(k) \, \log\big(p_m(k) + \varepsilon\big), \quad m \in \mathcal{M}, \tag{22}$$

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

$$\text{Rel}_m = \frac{\exp(-H_m)}{\sum_{m' \in \mathcal{M}} \exp(-H_{m'})}, \quad m \in \mathcal{M}. \tag{23}$$

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

$$u = W_{\text{imp}} z^{(0)} + b_{\text{imp}} \in \mathbb{R}^3, \tag{24}$$

$$\text{Imp}_m = \frac{\exp(u_m)}{\sum_{m' \in \mathcal{M}} \exp(u_{m'})}, \quad m \in \mathcal{M}. \tag{25}$$

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

$$\tilde{r}_m = \alpha \, \text{Rel}_m + (1 - \alpha) \, \text{Imp}_m, \quad m \in \mathcal{M}, \tag{26}$$

We normalize the relative modality weights and separately aggregate the unnormalized entropy-based confidence scores:

$$r_m = \frac{\tilde{r}_m}{\sum_{m' \in \mathcal{M}} \tilde{r}_{m'}}, \quad \bar{r} = \exp\left(-\frac{H_T + H_A + H_V}{3}\right). \tag{27}$$

The r_m are relative weights used in uniqueness regularization and expert-prior construction. In contrast, $\bar{r}$ is the geometric mean of $\exp(-H_T)$, $\exp(-H_A)$, and $\exp(-H_V)$; it is not the arithmetic mean of r_m. Consequently, $\bar{r}^3 = \exp(-(H_T + H_A + H_V))$ is a product of confidence proxies, not a probability that all three modalities are correct. The confidence weights remain differentiable in the objective; no stop-gradient is applied to r_m or $\bar{r}$. The consensus score defined next measures agreement separately.

B.1.5 CONSENSUS SCORE FROM UNIMODAL PREDICTIONS

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

$$\bar{p} = \frac{1}{3}\big(p_T + p_A + p_V\big). \tag{28}$$

For each modality,

$$D_m = \mathrm{KL}(p_m \,\|\, \bar{p}) = \sum_{k=1}^{K} p_m(k)\, \log \frac{p_m(k) + \varepsilon}{\bar{p}(k) + \varepsilon}, \tag{29}$$

and the mean disagreement is

$$\text{Disagree} = \frac{1}{3}\big(D_T + D_A + D_V\big). \tag{30}$$

A scalar consensus score is then defined as

$$\text{cons_score} = \exp\big(-\gamma \cdot \text{Disagree}\big), \quad \gamma > 0. \tag{31}$$

High cons_score indicates similar probe distributions, while lower values indicate greater disagreement. This signal complements entropy-based confidence and enters a learned mapping to expert priors; it does not prescribe a fixed expert choice or establish which modality is correct.

B.1.6 EXPERT PRIORS AND LOG-PRIOR INJECTION

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

$$f = \begin{bmatrix} \mathrm{Prior}_T^{\text{mod}} \\ \mathrm{Prior}_A^{\text{mod}} \\ \mathrm{Prior}_V^{\text{mod}} \\ \text{cons_score} \end{bmatrix} \in \mathbb{R}^4 \tag{32}$$

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

$$u^{\text{exp}} = W_{\text{prior}} f + b_{\text{prior}} \in \mathbb{R}^5, \tag{33}$$

which are converted into an expert prior distribution

$$\mathrm{Prior}_e^{\text{exp}} = \frac{\exp(u_e^{\text{exp}})}{\sum_{e' \in \mathcal{E}} \exp(u_{e'}^{\text{exp}})}, \quad e \in \mathcal{E}. \tag{34}$$

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

$$z_{\text{router}} = f_{\text{router}}\Big(z^{(0)}\Big) \in \mathbb{R}^5. \tag{35}$$

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

$$\tilde{z}_{\text{router},e} = z_{\text{router},e} + \lambda_p \log\big(\mathrm{Prior}_e^{\text{exp}} + \varepsilon\big), \quad e \in \mathcal{E}, \tag{36}$$

and obtain the final routing probabilities by softmax

$$\pi_e^{(0)} = \frac{\exp\big(\tilde{z}_{\text{router},e}\big)}{\sum_{e' \in \mathcal{E}} \exp\big(\tilde{z}_{\text{router},e'}\big)}. \tag{37}$$

Thus the router remains trained end-to-end through $\mathcal{L}_{\text{task}}$, 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 ($p = -T$):

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

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

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

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

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

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:

$$z^{(p)} = f_{\text{fuse}}\big(h_T^{(p)}, h_A^{(p)}, h_V^{(p)}\big), \tag{38}$$

$$y_e^{(p)} = f_e\Big(z^{(p)}\Big), \quad e \in \mathcal{E}. \tag{39}$$

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 expand the compact DIR formulas in Eqs. (3)–(8), matching the definitions 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:

$$\begin{aligned} P_T &= \{0, -A, -V\}, \\ P_A &= \{0, -T, -V\}, \\ P_V &= \{0, -T, -A\}. \end{aligned} \tag{40}$$

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

$$\mathcal{L}_{\text{uni},m} = \frac{1}{|P_m| - 1} \sum_{p \in P_m \setminus \{0\}} \big(y_{U_m}^{(0)} - y_{U_m}^{(p)}\big)^2. \tag{41}$$

The global uniqueness loss, weighted by modality reliability, is

$$\mathcal{L}_{\text{uni}} = \sum_{m \in \mathcal{M}} r_m\, \mathcal{L}_{\text{uni},m}. \tag{42}$$

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

$$\bar{y}_R = \frac{1}{4} \sum_{p \in \{0, -T, -A, -V\}} y_R^{(p)}, \tag{43}$$

and define

$$\mathcal{L}_{\text{red}} = \bar{r} \cdot \frac{1}{4} \sum_{p \in \{0, -T, -A, -V\}} \big(y_R^{(p)} - \bar{y}_R\big)^2. \tag{44}$$

This variance penalty encourages cross-view prediction stability. It does not selectively identify shared information, and predictive usefulness must be assessed together with the supervised objective and empirical results.

The PID-inspired invariance loss is

$$\mathcal{L}_{\text{PID}} = \mathcal{L}_{\text{uni}} + \mathcal{L}_{\text{red}}. \tag{45}$$

B.3.2 SYNERGY REGULARIZATION

The second DIR component, $\mathcal{L}_{\text{syn}}$, combines masked-view suppression with full-view residual fitting for the designated synergy pathway.

Synergy off under masking. We encourage the designated S pathway to have small outputs on the three single-modality masked views:

$$\mathcal{L}_{\text{syn-off}} = \bar{r}^3 \cdot \frac{1}{3} \sum_{p \in \{-T,-A,-V\}} \big(y_S^{(p)}\big)^2. \tag{46}$$

The factor $\bar{r}^3 = \exp(-(H_T + H_A + H_V))$ decreases as any probe entropy increases, with the other entropies fixed. It is a confidence-based weight and does not guarantee that high-entropy samples are corrupted or that low-entropy samples are correct.

Residual fit on the full view. On the full view, the routed synergy contribution is trained against the residual of the current non-synergy prediction, rather than information that the other experts are fundamentally unable to represent. Let

$$\hat{y}_{\setminus S}^{(0)} = \sum_{e \in \mathcal{E} \setminus \{S\}} \pi_e^{(0)}\, y_e^{(0)}, \tag{47}$$

and define a residual target with stop-gradient

$$\delta = y - \text{sg}\big(\hat{y}_{\setminus S}^{(0)}\big), \tag{48}$$

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

$$\mathcal{L}_{\text{syn-res}} = \bar{r}^3 \big(\pi_S^{(0)} y_S^{(0)} - \delta\big)^2. \tag{49}$$

Stop-gradient blocks the target branch through $\hat{y}_{\setminus S}^{(0)}$. Gradients still pass through $\pi_S^{(0)} y_S^{(0)}$ into the synergy head, the coupled softmax router, and shared upstream modules, and through $\bar{r}^3$ into the probes and their upstream modules. Thus this term gives no direct gradient to the non-synergy heads' private parameters, but it can change their subsequent predictions through shared parameters and routing. Appendix A gives the explicit gradient expression.

Total synergy loss. The synergy loss is

$$\mathcal{L}_{\text{syn}} = \mathcal{L}_{\text{syn-off}} + \mathcal{L}_{\text{syn-res}}. \tag{50}$$

Together, $\mathcal{L}_{\text{syn-off}}$ and $\mathcal{L}_{\text{syn-res}}$ encourage suppressed outputs on the specified masked views and a routed residual correction on the full view. These are predictive training roles rather than identification of an information-theoretic synergy component.

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). **CH-SIMS** is a Chinese single- and multi-modal sentiment analysis dataset containing 2,281 refined video segments collected from 60 raw videos. We follow the standard split with 1,368 samples for training, 456 for

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. (9) 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 $\hat{y}^{(0)} = \sum_e \pi_e^{(0)} y_e^{(0)}$.

validation, and 457 for testing. Each segment provides multimodal sentiment annotations as well as independent unimodal annotations for text, audio, and vision. The final sentiment labels lie in $[-1, 1]$ and are discretized into five categories: negative, weakly negative, neutral, weakly positive, and 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:

$$\mathrm{AUC}(M) = \sum_{k=1}^{K} (s_k - s_{k-1}) \cdot \frac{M(s_k) + M(s_{k-1})}{2}. \tag{51}$$

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

$$\mathrm{nAUC}(M) = \frac{\mathrm{AUC}(M)}{s_K - s_0}, \tag{52}$$

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:

$$\Delta U_m = \mathbb{E}_i \left[\frac{1}{|P_m| - 1} \sum_{p \in P_m \setminus \{0\}} \left| y_{U_m}^{(0)}(i) - y_{U_m}^{(p)}(i) \right| \right]. \tag{53}$$

and report $\Delta U_{\mathrm{avg}} = \frac{1}{3} \sum_{m \in \{T,A,V\}} \Delta U_m$. Lower ΔU indicates greater output stability under the specified removals of other modalities; it does not establish absence of cross-modal dependence.

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

$$\mathrm{Var}(R) \;=\; \mathbb{E}_i \left[\mathrm{Var}_{p \in \{0, -T, -A, -V\}} \left(y_R^{(p)}(i) \right) \right], \tag{54}$$

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

$$\mathrm{Off}(S) \;=\; \mathbb{E}_i \left[\frac{1}{3} \sum_{p \in \{-T, -A, -V\}} \left| y_S^{(p)}(i) \right| \right]. \tag{55}$$

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:

$$\mathrm{corr}\Big(\pi_S^{(0)} y_S^{(0)},\, \delta\Big) = \mathrm{PearsonCorr}\Big(\{\pi_S^{(0)}(i) y_S^{(0)}(i)\}_i, \{\delta(i)\}_i\Big). \tag{56}$$

Higher correlation indicates stronger alignment with the current non-synergy residual. It does not distinguish single-modality prediction error from genuinely joint information.

C.5 Implementation details

Input features. For CMU-MOSI and CMU-MOSEI, we use 300-dimensional GloVe language features (Pennington et al., 2014) and 768-dimensional BERT-base-uncased hidden states (Devlin et al., 2019) for the textual modality. The visual and acoustic modalities are represented by 35-dimensional Facet facial action unit features (Baltrušaitis et al., 2016) and 74-dimensional COVAREP acoustic features (Degottex et al., 2014), respectively. For CH-SIMS (Yu et al., 2020), we use the standard preprocessed features, including 768-dimensional Chinese BERT-base textual features, 33-dimensional LibROSA acoustic features, and 709-dimensional OpenFace2.0 visual features.

Training protocol. All reported hyperparameters are selected on the validation set, and the corresponding test performance is obtained under the same training configuration without test-set tuning. Experiments are implemented in PyTorch and conducted on a single NVIDIA A100 GPU with 40GB memory. Unless otherwise specified, all models are trained for 60 epochs with a batch size of 32 using the AdamW optimizer.

Dataset-specific settings. For CMU-MOSI, we use the compact-capacity configuration by setting the MoE hidden size to $d_{\text{model}} = 128$. Each Transformer-based expert has $L_e = 4$ 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. The reliability-aware routing hyperparameters are $\alpha = 0.7$, $\gamma = 1.0$, and $\lambda_p = 1.0$. 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}.

For CMU-MOSEI, we use a larger-capacity configuration by setting $d_{\text{model}} = 512$. Each Transformer-based expert has $L_e = 8$ layers, $H_e = 8$ attention heads, feed-forward dimension $d_{\text{ff}} = 1024$, and dropout rate 0.3. The probe and routing hyperparameters are kept the same as CMU-MOSI, i.e., $K = 7$, $\alpha = 0.7$, $\gamma = 1.0$, and $\lambda_p = 1.0$. 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}.

For CH-SIMS, we adopt the same compact-capacity configuration as CMU-MOSI, namely $d_{\text{model}} = 128$, $L_e = 4$, $H_e = 4$, $d_{\text{ff}} = 256$, and dropout rate 0.3. Since CH-SIMS uses a five-class sentiment label space, the probe classifier is changed to $K = 5$. The routing hyperparameters, loss weights, learning rate, and weight decay are kept the same as CMU-MOSI: $\alpha = 0.7$, $\gamma = 1.0$, $\lambda_p = 1.0$, $\lambda_{\text{cls}} = 0.05$, $\lambda_{\text{PID}} = 0.05$, $\lambda_{\text{syn}} = 0.05$, learning rate 3×10^{-4}, and weight decay 5×10^{-4}.

C.6 Hyperparameter Sensitivity

We examine whether RIDER-MoE relies on a narrow hyperparameter choice. Instead of sweeping generic optimization hyperparameters, we focus on the two mechanism-specific groups in our model: the router prior-injection strength λ_p and the DIR loss weight. Since λ_{PID} and λ_{syn} play complementary roles within the same masked-view role-shaping objective, we sweep them with a balanced weight $\lambda_D = \lambda_{\text{PID}} = \lambda_{\text{syn}}$.

Specifically, we vary $\lambda_p \in \{0, 0.5, 1.0, 2.0\}$ while fixing $\lambda_D = 0.05$, and vary $\lambda_D \in \{0, 0.01, 0.05, 0.10\}$ while fixing $\lambda_p = 1.0$. The default setting is $\lambda_p = 1.0, \lambda_D = 0.05$. All other hyperparameters, encoders, validation protocol, and stress-test grids are fixed to the main

Table 6: **Clean-set hyperparameter sensitivity.** A7, A2, A5, F1, MAE, and C denote Acc-7, Acc-2, Acc-5, weighted F1, MAE, and Pearson correlation, respectively. MOSI and MOSEI report Acc-7, while CH-SIMS reports Acc-5. The default setting is $\lambda_p = 1.0$ and $\lambda_D = \lambda_{\text{PID}} = \lambda_{\text{syn}} = 0.05$.

Setting	MOSI					MOSEI					CH-SIMS				
	A7	A2	F1	MAE	C	A7	A2	F1	MAE	C	A2	A5	F1	MAE	C
$\lambda_p = 0$	46.6	84.6	84.8	0.715	0.792	54.1	85.6	85.7	0.536	0.768	80.7	44.0	80.6	0.411	0.609
$\lambda_p = 0.5$	46.8	84.9	85.0	0.711	0.794	54.4	85.8	86.0	0.532	0.772	81.1	44.2	80.9	0.407	0.614
default	46.9	84.8	85.1	0.709	0.796	54.3	85.9	86.1	0.531	0.772	81.2	44.3	81.1	0.405	0.615
$\lambda_p = 2.0$	47.0	84.6	84.9	0.712	0.795	54.2	85.8	85.9	0.533	0.771	81.0	44.1	81.2	0.406	0.613
$\lambda_D = 0$	46.2	84.0	84.3	0.728	0.784	53.8	84.9	84.9	0.548	0.758	80.0	43.2	79.9	0.417	0.600
$\lambda_D = 0.01$	46.4	84.3	84.5	0.722	0.789	54.0	85.2	85.3	0.541	0.765	80.5	43.6	80.3	0.413	0.604
$\lambda_D = 0.10$	46.8	84.7	84.8	0.710	0.797	54.5	85.7	86.0	0.532	0.772	80.9	44.4	80.9	0.406	0.614

Table 7: **Robustness hyperparameter sensitivity.** Each cell reports Acc-2 nAUC/Corr nAUC. N, M, and Cf denote Noise, Missingness, and Cross-modal Conflict, respectively. Overall averages all nine dataset–stressor cells. Higher is better.

Setting	MOSI			MOSEI			CH-SIMS			Overall
	N	M	Cf	N	M	Cf	N	M	Cf	
$\lambda_p = 0$	78.6/0.699	75.1/0.660	80.1/0.721	79.7/0.683	76.3/0.648	80.9/0.704	73.4/0.541	70.2/0.517	73.9/0.545	76.5/0.635
$\lambda_p = 0.5$	79.9/0.714	76.5/0.676	80.6/0.729	81.2/0.702	77.8/0.668	81.7/0.715	74.6/0.556	71.5/0.533	74.7/0.557	77.6/0.650
default	80.7/0.722	77.3/0.684	81.0/0.733	81.6/0.706	78.4/0.672	82.0/0.718	75.2/0.565	72.0/0.540	75.0/0.560	**78.1/0.656**
$\lambda_p = 2.0$	80.8/0.724	76.9/0.681	80.5/0.728	81.4/0.705	78.1/0.670	81.5/0.713	75.0/0.563	71.5/0.536	74.4/0.555	77.8/0.653
$\lambda_D = 0$	77.4/0.685	72.9/0.632	79.2/0.708	78.5/0.675	73.5/0.627	80.2/0.693	72.6/0.536	68.0/0.499	73.0/0.531	75.0/0.621
$\lambda_D = 0.01$	78.7/0.700	74.4/0.657	79.0/0.711	79.3/0.687	75.4/0.646	80.0/0.700	73.2/0.541	69.5/0.511	73.5/0.541	75.9/0.633
$\lambda_D = 0.10$	80.5/0.720	77.1/0.682	80.7/0.731	81.7/0.707	78.0/0.669	81.8/0.716	74.8/0.560	71.8/0.538	74.7/0.558	77.9/0.653

experiments. The setting $\lambda_p = 0$ removes expert-prior logit injection, while $\lambda_D = 0$ removes both DIR terms.

Findings. Tables 6 and 7 show that RIDER-MoE is not tuned to a narrow hyperparameter point. First, changing λ_p has only a small effect on clean metrics, but removing prior injection reduces robust nAUC from 78.1/0.656 to 76.5/0.635. This indicates that expert-prior injection mainly affects reliability-aware routing under degraded or conflicting inputs, rather than acting as a clean-data fitting shortcut.

Second, DIR is more important than the prior-injection strength. Setting $\lambda_D = 0$ reduces the overall robust average to 75.0/0.621, with the largest drop under Missingness. This is consistent with the role of DIR: the five experts remain present when $\lambda_D = 0$, but the masked-view constraints no longer stabilize the intended uniqueness, redundancy, and synergy behaviors.

Third, moderate changes around the default remain close to the full model. Both $\lambda_p = 0.5$ and $\lambda_p = 2.0$ preserve strong clean performance and robust nAUC, and $\lambda_D = 0.10$ is also close to the default. The default setting yields the strongest overall robustness, but the nearby configurations form a stable plateau rather than a sharp optimum.

Takeaway. The sensitivity results support the same mechanism as the main ablations: λ_p provides a useful reliability/consensus prior for test-time routing, while DIR is the stronger source of broad robustness because it preserves behaviorally distinct U/R/S expert roles under masked-view interventions.

Table 8: **Full robustness curves on CMU-MOSI.** Acc-2 and Corr values are listed in increasing stressor severity order. Bold nAUC marks the best value within each stressor and metric.

Stressor	Method	Acc-2 values	Acc-2 nAUC	Corr values	Corr nAUC
Noise, $\rho = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$					
N	Self-MM	[84.6, 82.5, 80.3, 78.1, 75.8, 74.0]	79.2	[0.793, 0.759, 0.724, 0.685, 0.650, 0.621]	0.705
N	TMBL	[83.8, 81.6, 79.2, 76.7, 74.4, 72.4]	78.0	[0.762, 0.731, 0.697, 0.662, 0.628, 0.601]	0.680
N	EMOE	[85.2, 83.4, 81.2, 78.7, 76.7, 74.7]	80.0	[0.792, 0.763, 0.727, 0.688, 0.654, 0.623]	0.708
N	DLF	[85.1, 83.1, 80.9, 78.5, 76.3, 74.3]	79.7	[0.781, 0.752, 0.718, 0.682, 0.648, 0.619]	0.700
N	GCNet	[83.7, 81.9, 79.7, 77.4, 75.3, 73.6]	78.6	[0.755, 0.728, 0.697, 0.662, 0.632, 0.607]	0.680
N	CorrKD	[84.1, 82.3, 80.1, 77.9, 75.7, 74.0]	79.0	[0.772, 0.742, 0.708, 0.672, 0.637, 0.610]	0.690
N	LNLN	[84.7, 83.2, 81.4, 79.0, 77.0, 75.0]	80.1	[0.788, 0.762, 0.732, 0.693, 0.657, 0.624]	0.710
N	TMDC	[84.9, 83.6, 81.9, 79.9, 78.0, 76.3]	**80.8**	[0.790, 0.766, 0.736, 0.700, 0.666, 0.634]	0.716
N	**RIDER-MoE**	[84.8, 83.3, 81.6, 79.9, 78.1, 76.5]	80.7	[0.796, 0.769, 0.738, 0.707, 0.675, 0.646]	**0.722**
Missingness, $m = \{0, 0.2, 0.4, 0.6, 0.8\}$					
M	Self-MM	[84.6, 80.1, 74.7, 69.1, 64.4]	74.6	[0.793, 0.727, 0.646, 0.564, 0.494]	0.645
M	TMBL	[83.8, 79.3, 73.9, 68.3, 63.6]	73.8	[0.762, 0.702, 0.629, 0.555, 0.491]	0.628
M	EMOE	[85.2, 80.8, 75.6, 69.9, 64.6]	75.3	[0.792, 0.731, 0.656, 0.578, 0.503]	0.653
M	DLF	[85.1, 80.6, 75.3, 69.6, 64.8]	75.1	[0.781, 0.721, 0.651, 0.575, 0.510]	0.648
M	GCNet	[83.7, 80.3, 76.1, 71.8, 67.8]	76.0	[0.755, 0.710, 0.658, 0.600, 0.549]	0.655
M	CorrKD	[84.1, 80.8, 76.7, 72.5, 68.6]	76.6	[0.772, 0.724, 0.669, 0.606, 0.551]	0.665
M	LNLN	[84.7, 81.2, 77.2, 72.8, 69.0]	77.0	[0.788, 0.737, 0.677, 0.613, 0.558]	0.675
M	TMDC	[84.9, 81.6, 77.6, 73.2, 69.6]	**77.4**	[0.790, 0.744, 0.688, 0.628, 0.578]	**0.686**
M	**RIDER-MoE**	[84.8, 81.4, 77.4, 73.2, 69.6]	77.3	[0.796, 0.745, 0.686, 0.623, 0.569]	0.684
Cross-modal Conflict, $c = \{0, 0.25, 0.5, 0.75, 1.0\}$					
Cf	Self-MM	[84.6, 82.0, 78.6, 75.1, 71.1]	78.4	[0.793, 0.749, 0.695, 0.635, 0.569]	0.690
Cf	TMBL	[83.8, 81.0, 77.5, 73.7, 69.4]	77.2	[0.762, 0.719, 0.664, 0.606, 0.540]	0.660
Cf	EMOE	[85.2, 82.9, 80.2, 76.6, 73.0]	79.7	[0.792, 0.758, 0.717, 0.664, 0.610]	0.710
Cf	DLF	[85.1, 83.0, 80.5, 77.5, 74.5]	80.2	[0.781, 0.755, 0.721, 0.683, 0.644]	0.718
Cf	GCNet	[83.7, 80.7, 77.2, 73.1, 68.7]	76.8	[0.755, 0.716, 0.670, 0.617, 0.560]	0.665
Cf	CorrKD	[84.1, 81.2, 78.0, 73.9, 69.8]	77.5	[0.772, 0.728, 0.676, 0.615, 0.550]	0.670
Cf	LNLN	[84.7, 82.2, 79.0, 75.6, 72.0]	78.8	[0.788, 0.751, 0.704, 0.652, 0.598]	0.700
Cf	TMDC	[84.9, 82.0, 78.7, 74.7, 70.6]	78.3	[0.790, 0.746, 0.694, 0.633, 0.568]	0.688
Cf	**RIDER-MoE**	[84.8, 83.2, 81.2, 78.9, 76.6]	**81.0**	[0.796, 0.769, 0.738, 0.698, 0.659]	**0.733**

D Additional Robustness and Ablation Results

D.1 Full Robustness Curves under Test-Time Stressors

The main text reports normalized AUC (nAUC) summaries under Noise (N), Missingness (M), and Cross-modal Conflict (Cf). Here we provide the underlying performance–severity curves. For Noise, severities are $\rho = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$; for Missingness, $m = \{0, 0.2, 0.4, 0.6, 0.8\}$; for Conflict, $c = \{0, 0.25, 0.5, 0.75, 1.0\}$. Each row reports the full Acc-2 and Corr trajectory followed by its nAUC.

Across the full curves, RIDER-MoE is most consistently separated from competing methods under Cross-modal Conflict. This supports the interpretation in the main text: conflict is not merely a low-quality input regime, since unimodal signals may remain confident while their semantics disagree. Reliability- and consensus-conditioned routing therefore helps by moving probability mass away from fragile synergy and toward modality-specific evidence when agreement drops.

Table 9: **Full robustness curves on CMU-MOSEI.** Acc-2 and Corr values are listed in increasing stressor severity order. Bold nAUC marks the best value within each stressor and metric.

Stressor	Method	Acc-2 values	Acc-2 nAUC	Corr values	Corr nAUC
Noise, $\rho = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$					
N	Self-MM	[85.0, 83.1, 81.0, 78.7, 76.7, 75.0]	79.9	[0.766, 0.730, 0.692, 0.651, 0.613, 0.582]	0.672
N	TMBL	[85.8, 83.9, 81.8, 79.5, 77.5, 75.8]	80.7	[0.766, 0.731, 0.693, 0.654, 0.616, 0.586]	0.674
N	EMOE	[85.1, 83.4, 81.4, 79.2, 77.2, 75.4]	80.3	[0.759, 0.732, 0.699, 0.665, 0.633, 0.604]	0.682
N	DLF	[85.4, 83.8, 82.0, 80.0, 78.2, 76.6]	81.0	[0.764, 0.737, 0.705, 0.671, 0.639, 0.612]	0.688
N	GCNet	[84.6, 82.9, 80.9, 78.6, 76.7, 75.1]	79.8	[0.742, 0.715, 0.684, 0.652, 0.620, 0.596]	0.668
N	CorrKD	[85.0, 83.4, 81.5, 79.4, 77.6, 76.1]	80.5	[0.755, 0.727, 0.695, 0.661, 0.628, 0.603]	0.678
N	LNLN	[85.3, 84.0, 82.4, 80.4, 78.6, 76.9]	81.3	[0.766, 0.743, 0.715, 0.680, 0.649, 0.619]	0.696
N	TMDC	[85.6, 84.3, 82.8, 80.8, 79.1, 77.4]	**81.7**	[0.770, 0.748, 0.722, 0.690, 0.659, 0.631]	0.704
N	**RIDER-MoE**	[85.9, 84.3, 82.5, 80.7, 78.9, 77.2]	81.6	[0.772, 0.748, 0.720, 0.693, 0.664, 0.638]	**0.706**
Missingness, $m = \{0, 0.2, 0.4, 0.6, 0.8\}$					
M	Self-MM	[85.0, 80.6, 75.1, 69.7, 65.0]	75.1	[0.766, 0.700, 0.619, 0.537, 0.467]	0.618
M	TMBL	[85.8, 81.4, 76.0, 70.7, 66.0]	76.0	[0.766, 0.701, 0.621, 0.540, 0.471]	0.620
M	EMOE	[85.1, 81.0, 75.9, 70.8, 65.8]	75.8	[0.759, 0.703, 0.635, 0.563, 0.495]	0.632
M	DLF	[85.4, 81.2, 76.2, 71.0, 66.5]	76.1	[0.764, 0.708, 0.639, 0.569, 0.508]	0.638
M	GCNet	[84.6, 81.2, 77.0, 72.7, 68.7]	76.9	[0.742, 0.700, 0.651, 0.596, 0.548]	0.648
M	CorrKD	[85.0, 81.7, 77.6, 73.4, 69.5]	77.5	[0.755, 0.711, 0.660, 0.603, 0.553]	0.657
M	LNLN	[85.3, 82.1, 78.3, 74.3, 70.9]	78.2	[0.766, 0.722, 0.669, 0.614, 0.567]	0.668
M	TMDC	[85.6, 82.5, 78.6, 74.5, 71.2]	**78.5**	[0.770, 0.728, 0.676, 0.620, 0.575]	**0.674**
M	**RIDER-MoE**	[85.9, 82.5, 78.5, 74.3, 70.7]	78.4	[0.772, 0.726, 0.674, 0.618, 0.569]	0.672
Cross-modal Conflict, $c = \{0, 0.25, 0.5, 0.75, 1.0\}$					
Cf	Self-MM	[85.0, 82.3, 79.1, 75.3, 71.3]	78.7	[0.766, 0.721, 0.664, 0.604, 0.535]	0.660
Cf	TMBL	[85.8, 82.9, 79.3, 75.4, 71.0]	79.0	[0.766, 0.722, 0.667, 0.607, 0.539]	0.662
Cf	EMOE	[85.1, 83.0, 80.3, 77.2, 73.8]	80.0	[0.759, 0.731, 0.695, 0.651, 0.606]	0.690
Cf	DLF	[85.4, 83.5, 81.1, 78.2, 75.4]	80.8	[0.764, 0.739, 0.707, 0.671, 0.634]	0.704
Cf	GCNet	[84.6, 81.8, 78.5, 74.8, 70.7]	78.2	[0.742, 0.704, 0.661, 0.608, 0.553]	0.655
Cf	CorrKD	[85.0, 82.4, 79.2, 75.6, 71.8]	78.9	[0.755, 0.716, 0.670, 0.617, 0.560]	0.665
Cf	LNLN	[85.3, 83.1, 80.3, 77.3, 74.1]	80.1	[0.766, 0.734, 0.693, 0.649, 0.602]	0.690
Cf	TMDC	[85.6, 83.0, 80.1, 76.5, 72.8]	79.7	[0.770, 0.732, 0.687, 0.635, 0.579]	0.682
Cf	**RIDER-MoE**	[85.9, 84.2, 82.2, 79.9, 77.4]	**82.0**	[0.772, 0.749, 0.721, 0.688, 0.655]	**0.718**

Table 10: **Full robustness curves on CH-SIMS.** Acc-2 and Corr values are listed in increasing stressor severity order. Bold nAUC marks the best value within each stressor and metric.

Stressor	Method	Acc-2 values	Acc-2 nAUC	Corr values	Corr nAUC
Noise, $\rho = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$					
N	Self-MM	[78.6, 76.2, 73.7, 71.1, 68.5, 66.5]	72.4	[0.601, 0.576, 0.549, 0.519, 0.494, 0.472]	0.535
N	TMBL	[79.1, 76.7, 74.2, 71.6, 69.0, 67.0]	72.9	[0.592, 0.566, 0.539, 0.511, 0.483, 0.461]	0.525
N	EMOE	[75.7, 73.6, 70.9, 68.0, 65.5, 63.2]	69.5	[0.519, 0.497, 0.469, 0.441, 0.414, 0.390]	0.455
N	DLF	[80.1, 77.9, 75.4, 72.8, 70.3, 68.1]	74.1	[0.563, 0.544, 0.522, 0.498, 0.476, 0.457]	0.510
N	GCNet	[76.4, 74.4, 72.2, 69.9, 67.5, 65.7]	71.0	[0.525, 0.516, 0.506, 0.494, 0.484, 0.476]	0.500
N	CorrKD	[77.2, 75.2, 73.0, 70.6, 68.3, 66.5]	71.8	[0.545, 0.533, 0.519, 0.505, 0.491, 0.480]	0.512
N	LNLN	[79.2, 77.8, 76.2, 74.0, 72.2, 70.4]	75.0	[0.575, 0.567, 0.558, 0.548, 0.536, 0.527]	0.552
N	TMDC	[78.9, 77.5, 75.7, 73.6, 71.5, 69.6]	74.5	[0.570, 0.562, 0.552, 0.539, 0.528, 0.517]	0.545
N	**RIDER-MoE**	[81.2, 79.0, 76.5, 73.9, 71.4, 69.1]	**75.2**	[0.615, 0.597, 0.576, 0.554, 0.533, 0.514]	**0.565**
Missingness, $m = \{0, 0.2, 0.4, 0.6, 0.8\}$					
M	Self-MM	[78.6, 73.8, 67.8, 61.9, 56.8]	67.8	[0.601, 0.556, 0.500, 0.445, 0.397]	0.500
M	TMBL	[79.1, 74.2, 68.1, 62.1, 56.9]	68.1	[0.592, 0.547, 0.493, 0.437, 0.390]	0.492
M	EMOE	[75.7, 71.2, 65.7, 60.0, 54.5]	65.5	[0.519, 0.475, 0.423, 0.366, 0.313]	0.420
M	DLF	[80.1, 75.2, 69.3, 63.3, 57.9]	69.2	[0.563, 0.526, 0.480, 0.435, 0.394]	0.480
M	GCNet	[76.4, 72.9, 68.6, 64.2, 60.1]	68.5	[0.525, 0.507, 0.486, 0.463, 0.443]	0.485
M	CorrKD	[77.2, 73.5, 69.3, 64.5, 60.3]	69.0	[0.545, 0.521, 0.491, 0.460, 0.432]	0.490
M	LNLN	[79.2, 75.5, 71.2, 66.5, 62.5]	71.0	[0.575, 0.553, 0.527, 0.499, 0.475]	0.526
M	TMDC	[78.9, 75.3, 70.9, 66.1, 62.2]	70.7	[0.570, 0.548, 0.521, 0.492, 0.468]	0.520
M	**RIDER-MoE**	[81.2, 77.0, 72.2, 67.0, 62.5]	**72.0**	[0.615, 0.581, 0.541, 0.499, 0.463]	**0.540**
Cross-modal Conflict, $c = \{0, 0.25, 0.5, 0.75, 1.0\}$					
Cf	Self-MM	[78.6, 75.2, 70.8, 66.2, 61.0]	70.5	[0.601, 0.567, 0.523, 0.477, 0.425]	0.520
Cf	TMBL	[79.1, 75.6, 71.1, 66.4, 61.0]	70.8	[0.592, 0.556, 0.512, 0.463, 0.409]	0.508
Cf	EMOE	[75.7, 72.7, 69.0, 64.3, 59.5]	68.4	[0.519, 0.486, 0.447, 0.396, 0.343]	0.440
Cf	DLF	[80.1, 77.2, 73.6, 69.4, 65.1]	73.2	[0.563, 0.542, 0.515, 0.484, 0.452]	0.512
Cf	GCNet	[76.4, 73.1, 69.4, 64.9, 60.1]	68.9	[0.525, 0.511, 0.494, 0.474, 0.453]	0.492
Cf	CorrKD	[77.2, 73.8, 69.8, 65.0, 60.0]	69.3	[0.545, 0.525, 0.503, 0.476, 0.447]	0.500
Cf	LNLN	[79.2, 76.8, 73.9, 70.5, 67.1]	73.6	[0.575, 0.561, 0.544, 0.524, 0.504]	0.542
Cf	TMDC	[78.9, 75.9, 72.6, 68.4, 64.1]	72.1	[0.570, 0.550, 0.528, 0.501, 0.472]	0.525
Cf	**RIDER-MoE**	[81.2, 78.5, 75.4, 71.6, 67.7]	**75.0**	[0.615, 0.591, 0.564, 0.530, 0.496]	**0.560**

D.2 Per-Dataset Component Ablations

The main text reports macro-averaged ablations over the three benchmarks. Tables 11–13 provide the corresponding per-dataset results. Each cell reports Acc-2/Corr. Clean denotes the uncorrupted test set, while Noise, Missingness, and Conflict report nAUC over the corresponding severity grid. Avg averages the three stressors within each dataset.

The dataset-level ablations show the same pattern as the macro-averaged results. Removing the expert prior reduces robustness but preserves most clean accuracy, indicating that prior injection mainly affects routing under stress. Removing the consensus feature has a targeted effect on Conflict, especially on CMU-MOSI and CH-SIMS. Removing DIR produces a larger and broader degradation, with the largest losses under Missingness, confirming that masked-view role shaping is central to preserving distinct U/R/S behavior. Finally, the capacity-matched single-head variant performs worst on every dataset-level robust average, ruling out additional head capacity as the primary explanation.

Table 11: **Per-dataset ablation on CMU-MOSI.** Each cell reports Acc-2/Corr; higher is better.

Variant	Clean	Noise	Missing	Conflict	Avg
RIDER-MoE	**84.8/0.796**	**80.7/0.722**	**77.3/0.684**	**81.0/0.733**	**79.7/0.713**
w/o prior injection ($\lambda_p = 0$)	84.6/0.792	78.6/0.699	75.1/0.660	80.1/0.721	77.9/0.693
w/o consensus feature	84.7/0.793	80.3/0.716	77.1/0.680	78.7/0.704	78.7/0.700
w/o DIR	84.0/0.784	77.4/0.685	72.9/0.632	79.2/0.708	76.5/0.675
w/o $\mathcal{L}_{\mathrm{PID}}$	84.4/0.789	80.0/0.713	76.0/0.672	79.1/0.714	78.4/0.700
w/o $\mathcal{L}_{\mathrm{syn}}$	84.5/0.790	79.1/0.704	75.2/0.660	79.8/0.716	78.0/0.693
single-head, capacity-matched	83.4/0.776	75.8/0.668	70.8/0.609	78.0/0.691	74.9/0.656

Table 12: **Per-dataset ablation on CMU-MOSEI.** Each cell reports Acc-2/Corr; higher is better.

Variant	Clean	Noise	Missing	Conflict	Avg
RIDER-MoE	**85.9/0.772**	**81.6/0.706**	78.4/**0.672**	**82.0/0.718**	**80.7/0.699**
w/o prior injection ($\lambda_p = 0$)	85.6/0.768	79.7/0.683	76.3/0.648	80.9/0.704	79.0/0.678
w/o consensus feature	85.7/0.770	81.1/0.700	**78.5**/0.671	79.3/0.688	79.6/0.686
w/o DIR	84.9/0.758	78.5/0.675	73.5/0.627	80.2/0.693	77.4/0.665
w/o $\mathcal{L}_{\mathrm{PID}}$	85.2/0.764	80.8/0.698	76.7/0.660	80.1/0.700	79.2/0.686
w/o $\mathcal{L}_{\mathrm{syn}}$	85.2/0.765	80.0/0.688	75.7/0.647	80.8/0.704	78.8/0.680
single-head, capacity-matched	84.0/0.748	76.0/0.656	71.1/0.603	78.5/0.685	75.2/0.648

Table 13: **Per-dataset ablation on CH-SIMS.** Each cell reports Acc-2/Corr; higher is better.

Variant	Clean	Noise	Missing	Conflict	Avg
RIDER-MoE	**81.2/0.615**	**75.2/0.565**	**72.0/0.540**	**75.0/0.560**	**74.1/0.555**
w/o prior injection ($\lambda_p = 0$)	80.7/0.609	73.4/0.541	70.2/0.517	73.9/0.545	72.5/0.534
w/o consensus feature	81.0/0.611	74.8/0.557	71.7/0.535	72.3/0.530	72.9/0.541
w/o DIR	80.0/0.600	72.6/0.536	68.0/0.499	73.0/0.531	71.2/0.522
w/o $\mathcal{L}_{\mathrm{PID}}$	80.5/0.606	74.4/0.552	70.8/0.525	73.5/0.540	72.9/0.539
w/o $\mathcal{L}_{\mathrm{syn}}$	80.4/0.606	73.6/0.544	70.2/0.518	73.6/0.544	72.5/0.535
single-head, capacity-matched	78.9/0.579	70.4/0.509	66.3/0.471	71.2/0.514	69.3/0.498