QA-MoE: Towards a Continuous Reliability Spectrum with Quality-Aware Mixture of Experts for Robust Multimodal Sentiment Analysis

Yitong Zhu[1] Yuxuan Jiang[2] Guanxuan Jiang[1] Bojing Hou[1]
Peng Yuan Zhou[3] Ge Lin KAN[1] Yuyang Wang[1]*
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Tsinghua University [3]Aarhus University
{yzhu162, gjiang240, bhou870}@connect.hkust-gz.edu.cn
jiangyux25@mails.tsinghua.edu.cn
{gelin,yuyangwang}@hkust-gz.edu.cn pengyuan.zhou@ece.au.dk

Abstract

Multimodal Sentiment Analysis (MSA) aims to infer human sentiment from textual, acoustic, and visual signals. In real-world scenarios, however, multimodal inputs are often compromised by dynamic noise or modality missingness. Existing methods typically treat these imperfections as discrete cases or assume fixed corruption ratios, which limits their adaptability to continuously varying reliability conditions. To address this, we first introduce a Continuous Reliability Spectrum to unify missingness and quality degradation into a single framework. Building on this, we propose QA-MoE, a Quality-Aware Mixture-of-Experts framework that quantifies modality reliability via self-supervised aleatoric uncertainty. This mechanism explicitly guides expert routing, enabling the model to suppress error propagation from unreliable signals while preserving task-relevant information. Extensive experiments indicate that QA-MoE achieves competitive or state-of-the-art performance across diverse degradation scenarios and exhibits a promising One-Checkpoint-for-All property in practice.

1 Introduction

Multimodal Sentiment Analysis (MSA) is a base of human-centric computing, which aims to explain complex emotional states by integrating Textual (T), Vision (V), and Acoustic (A) (Sun and Tian, 2025; Yang et al., 2024; Zhang et al., 2025; He et al., 2025b; Jiang et al., 2026). With recent advances in multimodal learning (Chen et al., 2023; Mizrahi et al., 2023; Zhu et al., 2024), MSA models (Fang et al., 2025; He et al., 2025a) can better exploit the complementary signals across modalities and capture subtle affective cues that unimodal systems often miss, narrowing the gap between human expression and machine understanding.

However, unlike the clean and complete data found in laboratory settings, real-world multimodal signals are often noisy or incomplete. Most existing models rely on the assumption of ideal inputs, creating a significant gap between constrained training conditions and the complexities of practical application (Zhao et al., 2021; Xu et al., 2024). In practice, modality noise fluctuates dynamically due to environmental interference, while data incompleteness frequently arises from sensor failures. These uncertainties often manifest as a stochastic mixture, where noise and missingness co-occur non-uniformly across samples (Figure. 1). Crucially, these data defects are not discrete categories, but instead exist across a broad range of intensities, ranging from subtle noise to the total loss of signal.

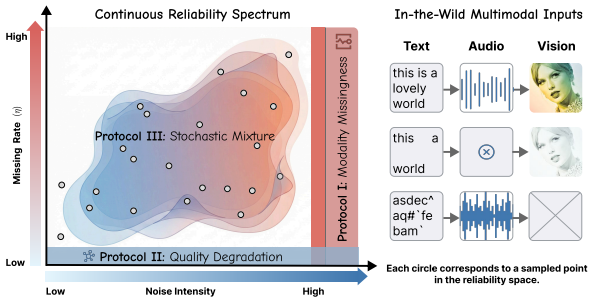

Figure 1: The Continuous Reliability Spectrum unifies three evaluation protocols defined by noise intensity (λ) and missing rate (η), and inputs from Text, Audio, and Vision are processed as imperfect multimodal inputs.

To tackle these reliability issues, earlier efforts have primarily focused on explicit data imputation, employing reconstruction methods (Cai et al., 2018; Lian et al., 2023; Guo et al., 2024) to recover missing modalities from the remaining observed signals. Recent advancements have pivoted toward architectural robustness, leveraging Bayesian meta-learning (Ma et al., 2021), diffusion models (Wang et al., 2023), and attention mechanisms (Mai et al., 2025) to directly learn from imperfect inputs. Nevertheless, these approaches still suffer from two critical limitations: (1) *Quality-Agnostic Extraction.* Existing models typically derive representa-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31689–31703

tions from raw inputs without explicitly modeling their reliability. Consequently, they fail to disentangle task-relevant semantics from non-informative noise, causing the model to capture spurious noisy artifacts rather than robust affective cues. (2) *Fixed-Ratio Bias.* These methods are optimized for specific, predefined corruption ratios seen during training. This rigidity prevents them from adapting to the fluctuating noise intensities in the wild, leading to severe performance drops when test-time reliability deviates from training protocols.

To bridge these gaps, we propose a unified framework centered around a **Continuous Reliability Spectrum** (Figure. 1). Rather than treating data imperfections as discrete cases, we conceptually map diverse defects onto this continuous spectrum, unifying three distinct evaluation protocols: *Modality Missingness*, *Quality Degradation*, and *Stochastic Mixture*. This perspective allows us to evaluate model robustness in a more holistic and realistic manner. Building upon this unified view, we introduce the **Quality-Aware Mixture of Experts** (*QA-MoE*). To move beyond traditional semantic-only routing, we incorporate a self-supervised reliability quantification module that utilizes aleatoric uncertainty to generate dynamic quality scores. These scores explicitly guide the MoE computation, transforming the routing process into a quality-aware aggregation. By weighting semantic gating with these reliability metrics, the framework effectively suppresses expert activation for unreliable inputs while prioritizing task-relevant signals. Finally, extensive experiments on MSA and Multimodal Intent Recognition (MIR) tasks demonstrate that QA-MoE achieves superior performance, validating the robustness and versatility of the proposed framework. Notably, evaluations across the comprehensive settings of the reliability spectrum reveal that our model effectively navigates varying levels of noise and missingness, establishing a **One-Checkpoint-for-All** capability. This signifies that a single trained model can generalize to arbitrary, unseen degradation intensities without retraining or specialized fine-tuning. The main contributions of our work are summarized as follows:

- We propose the Continuous Reliability Spectrum to unify modality missingness and quality degradation into a framework, moving beyond traditional treatment of discrete defects.
- We introduce a quality-aware mixture-of-experts framework, QA-MoE, in which self-supervised reliability estimation derived from aleatoric uncertainty explicitly guides routing to adapt expert aggregation to input quality.
- Extensive experiments on MSA benchmarks (CMU-MOSI, CMU-MOSEI) and cross-task datasets (IEMOCAP, MIntRec) show that our model achieves state-of-the-art performance, and exhibits a One-Checkpoint-for-All capability, generalizing across a wide range of noise levels and missing rates.

2 Related Work

2.1 Multimodal Sentiment Analysis

MSA integrates heterogeneous signals from language, vision, and acoustics to infer human emotions. Early work modeled explicit cross-modal interactions via tensor fusion (Zadeh et al., 2017). Transformer-based architectures (Tsai et al., 2019; Zhang et al., 2023) later advanced the field through cross-modal attention for aligning asynchronous streams. More recently, representation learning approaches (Zhu et al., 2025a) have emphasized disentanglement (Hazarika et al., 2020; Zhu et al., 2025b) and self-supervised objectives (Yang et al., 2024; He et al., 2025b) to reduce redundancy. However, most methods still assume complete and noise-free modalities. In contrast, we develop a unified framework that explicitly estimates signal reliability and operates across a continuous spectrum of degradation.

2.2 Imperfect Multimodal Learning

Real-world deployments often involve missing or noisy modalities. Early work addressed these imperfections through data imputation (Cai et al., 2018; Ma et al., 2021), reconstructing missing views via generative models; however, such methods are computationally expensive and vulnerable to distribution shift–induced hallucination (Guo et al., 2024). Subsequent research shifted toward robust representation learning, down-weighting corrupted features via attention (Mai et al., 2025) or auxiliary objectives (Wang et al., 2023), yet these approaches still struggle under severe noise. More recently, Mixture-of-Experts or Adapters have been adopted for multimodal learning (Xu et al., 2024; Chen et al., 2025), but existing routers are purely semantics-driven and fail to distinguish informative signals from corruption. To address this, we introduce a self-supervised quality signal into the

routing process in order to enable effective isolation of unreliable modalities.

3 Methodology

3.1 Preliminaries

To bridge the gap between idealized laboratory benchmarks and unpredictable real-world scenarios, we establish a foundational framework from standard encoding to reliability analysis.

Standard Multimodal Encoding. Given a multimodal dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, each sample X_i comprises Textual (t), Audio (a), and Video (v) modalities. Following standard protocols (Tsai et al., 2019), we employ pre-trained unimodal encoders $E_{m\in\{t,a,v\}}$ to extract the raw feature sequences $\mathbf{u}_m \in \mathbb{R}^{T_m \times d_m}$. In conventional lab-controlled settings, these extracted features are implicitly assumed to be complete and pristine.

From Ideal to Real. However, to capture the dynamic nature of real-world noise, we depart from this idealization. We first introduce the *Continuous Reliability Spectrum* (Sec. 3.2) to conceptually unify diverse data defects. Subsequently, we formalize this concept through *Stochastic Imperfection Modeling* (Sec. 3.3), which provides the theoretical basis for our evaluation protocols.

3.2 Continuous Reliability Spectrum

Considering that noise and missingness often occur simultaneously in real-world scenarios, we propose a unified reliability spectrum shown in Figure 1(left) to conceptually map these diverse imperfections onto a single latent measure. Instead of treating defects as discrete categories, we quantify the input quality by a **Latent Reliability Score** $r_m \in (0, 1]$:

$$\text{Degradation} \propto 1 - r_m \tag{1}$$

We define three characteristic phases along this spectrum based on r_m: **High Quality** ($r_m \approx 1$): Ideal clean data found in lab settings. **Quality Degradation** ($r_m \in (0, 1)$): Data corrupted by noise intensity λ_m. **Availability Limit** ($r_m \to 0$): Data subject to modality missingness rate η.

3.3 Stochastic Imperfection Modeling

Based on the unified spectrum, we formulate the real-world environment not as a static dataset, but as a **Stochastic Degradation Process**. For any input sample $\mathbf{u}_m$, the imperfect representation $\tilde{\mathbf{u}}_m$ is generated via a transformation function $\mathcal{T}$:

$$\tilde{\mathbf{u}}_m = (1 - \mathbb{I}_{\text{miss}}) \cdot (\mathbf{u}_m + \boldsymbol{\epsilon}_m) \tag{2}$$

$\mathbb{I}_{\text{miss}} \sim \text{Bernoulli}(\eta)$ here is a binary variable indicating modality absence. The term $\boldsymbol{\epsilon}_m \sim \mathcal{N}(\mathbf{0}, \sigma^2(\lambda_m)\mathbf{I})$ represents the additive noise, where the variance is governed by the degradation intensity λ_m.This formulation unifies three distinct evaluation protocols representing different subspaces of the reliability spectrum:

Protocol I: Modality Missingness. We focus on binary availability by setting $\boldsymbol{\epsilon}_m = \mathbf{0}$ and varying the missing rate η to simulate scenarios such as sensor failure or packet loss.

Protocol II: Quality Degradation. We focus on signal fidelity by fixing $\mathbb{I}_{\text{miss}} = 0$ and varying the noise intensity λ_m. This simulates noisy environments where the modality is present but unreliable.

Protocol III: Stochastic Mixture. We sample both λ_m and η from a joint distribution $P_{env}(\lambda, \eta)$. In this setting, input signals are subject to random combinations of noise corruption and modality unavailability, reflecting complex in-the-wild dynamics.

3.4 Quality-Aware Mixture of Experts

To effectively navigate the continuous reliability spectrum, we introduce the QA-MoE framework. Unlike standard deterministic models (Shi and Zhang, 2025; Zhang et al., 2024), QA-MoE operates under a probabilistic principle: it decouples the input representation into a semantic signal and an uncertainty measure, which is used to explicitly guide the computation flow.

3.4.1 Probabilistic Feature Modeling

Under our stochastic degradation process (Sec. 3.3), $\mathbf{u}_m$ is inevitably influenced by varying degrees of noise. To model the inherent uncertainty, we project the feature space onto a multivariate Gaussian distribution:

$$p(\mathbf{z}_m|\mathbf{u}_m) = \mathcal{N}(\mathbf{z}_m; \boldsymbol{\mu}_m, \text{diag}(\boldsymbol{\sigma}_m^2)) \tag{3}$$

, where $\mathbf{z}_m$ is the latent representation. We employ two parallel affine transformations to estimate the distribution parameters:

$$\boldsymbol{\mu}_m = \mathbf{W}_\mu \mathbf{u}_m + \mathbf{b}_\mu \tag{4}$$

$$\boldsymbol{\sigma}_m^2 = \text{Softplus}(\mathbf{W}_\sigma \mathbf{u}_m + \mathbf{b}_\sigma) \tag{5}$$

, $\mathbf{W}_{\mu/\sigma} \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_{\mu/\sigma} \in \mathbb{R}^d$ here are learnable projection matrices and bias terms. Specifically,

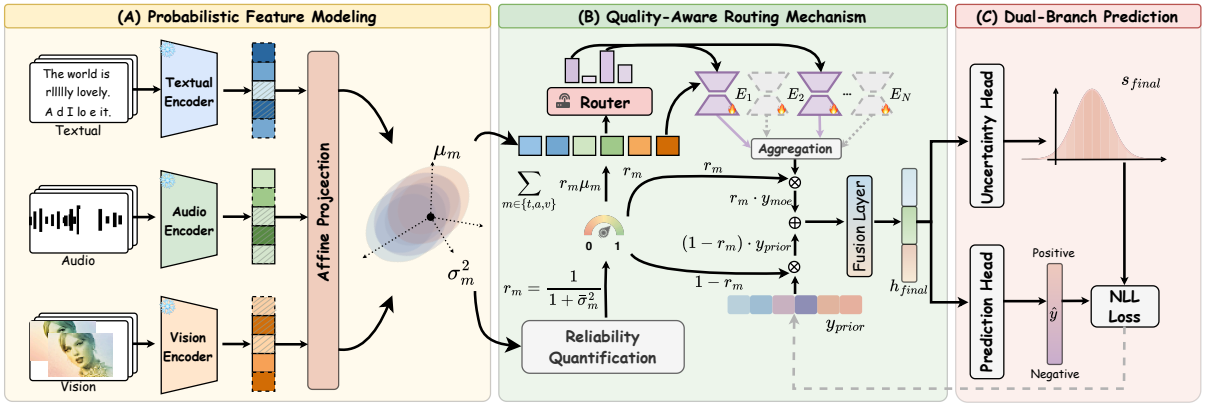

Figure 2: Overview of QA-MoE. (A) Probabilistic Feature Modeling encodes inputs as distributions to capture the uncertainty. (B) Quality-Aware Routing calculates a quality score r_m to guide the expert selection. (C) Dual-Branch Prediction outputs both the prediction and uncertainty to optimize the model via heteroscedastic regression.

$\boldsymbol{\mu}_m$ extracts the clean semantic signal, while $\boldsymbol{\sigma}_m^2$ quantifies the inherent uncertainty. High variance values indicate unreliable or missing inputs, serving as a dynamic quality indicator to guide the subsequent routing.

3.4.2 Quality-Aware Routing Mechanism

To solve the limitations of static computation, we design a quality-aware adaptive routing mechanism to dynamically align expert contribution with input reliability. It involves two key steps: quality quantification and modulating the expert aggregation.

Quality Quantification. First, we derive a scalar quality score $r_m \in (0, 1]$ to explicitly quantify the reliability of the input modality. Since the variance $\boldsymbol{\sigma}_m^2$ serves as an indicator of uncertainty, the quality is naturally modeled as inversely proportional to the aggregate variance:

$$r_m = \frac{1}{1 + \frac{1}{d}\sum_{k=1}^{d} \boldsymbol{\sigma}_{m,k}^2} \tag{6}$$

This formulation creates a bounded metric: when the input is clean, $r_m \to 1$; conversely, as degradation intensifies and variance explodes, r_m asymptotically decays to 0.

Selective Expert Aggregation. Next, we integrate the score into the MoE computation. We employ a bank of N experts $\{E_i\}_{i=1}^{N}$, where each expert is instantiated as a GLU to capture complex semantic patterns. A routing network computes the gating weights $g(\boldsymbol{\mu}_m) = \text{Softmax}(\mathbf{W}_g \boldsymbol{\mu}_m)$ based on the semantic centroid.

Crucially, we introduce r_m as a global suppression coefficient and the output $\mathbf{y}_m$ is computed via a smooth interpolation:

$$\mathbf{y}_m = r_m \cdot \sum_{i=1}^{N} g_i(\boldsymbol{\mu}_m) E_i(\boldsymbol{\mu}_m) + (1 - r_m) \cdot \mathbf{y}_{prior} \tag{7}$$

Here, $\mathbf{y}_{prior}$ is a learnable global static embedding, which captures a dataset-level semantic consensus independent of specific inputs. This allows the model to smoothly interpolate between the instance-specific prediction and this stable reference to mitigate the risk of overfitting to noise.

3.4.3 Dual-Branch Prediction

After obtaining the refined expert outputs $\mathbf{y}_m$ for all modalities, we aggregate them into a unified multimodal representation $\mathbf{h}_{final}$ via a standard fusion layer.

To enable the heteroscedastic regression objective (detailed in Sec. 3.5), the model must output not only a sentiment score but also a confidence measure. Therefore, we design a dual-branch regression head that decouples the prediction of value and total uncertainty:

$$\hat{y} = \mathbf{W}_y \mathbf{h}_{final} + b_y \tag{8}$$

$$\mathbf{s}_{final} = \mathbf{W}_s \mathbf{h}_{final} + b_s \tag{9}$$

Here, $\hat{y}$ and $\mathbf{s}_{final}$ is the predicted sentiment score and log-variance. $\mathbf{s}_{final}$ acts as a learned estimator of the total prediction uncertainty for the current sample, which is used to dynamically weigh the gradient updates during optimization.

3.5 Training and Optimization

To empower QA-MoE with the capability to generalize across the continuous reliability spectrum

in Sec. 1, we propose a unified learning framework that integrates a dynamic data augmentation strategy with a uncertainty-aware objective function.

Spectrum-Aware Training Strategy. Instead of relying on perfect datasets, we construct a dynamic training process to simulate real-world imperfections according to Sec 3.3. Specifically, for each training batch, we randomly inject noise and mask modalities with varying probabilities. This exposes the router to the full reliability spectrum during optimization. Detailed protocols for Spectrum Dataset generation are provided in Appendix A.1.

Optimization Objectives. To effectively train the QA-MoE framework under the Stochastic Degradation Protocol, we treat the model's output as a Gaussian distribution rather than a deterministic point estimate. $\hat{y}$ and $\mathbf{s}_{final}$ derived in Sec. 3.4.3 are used to minimize the Negative Log-Likelihood (NLL) of the ground truth y. The total loss $\mathcal{L}$ establishes a self-supervised feedback loop, which is formulated as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}e^{-\mathbf{s}_{final,i}}(y_i - \hat{y}_i)^2 + \frac{1}{2}\mathbf{s}_{final,i}\right) \tag{10}$$

For noisy inputs with large errors, reducing $\mathcal{L}$ forces $\mathbf{s}_{final}$ to increase. The gradient backpropagates to elevate the variance $\boldsymbol{\sigma}^2$, which directly reduces the quality score r_m. Consequently, the router learns to suppress experts for degraded data without manual labels.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets and Feature Extraction

To evaluate the robustness of our framework, we conduct experiments on four benchmarks: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018), IEMOCAP (Busso et al., 2008), and MIntRec (Zhang et al., 2022). Regarding feature extraction, we strictly adhere to the standard protocols from prior literature (Tsai et al., 2019; Hazarika et al., 2020). Detailed description, statistics and other information are provided in Appendix A.2.

4.1.2 Implementation Details

We give brief information about the baselines, metrics and the implementation. And the detailed description is provided in Appendix A.3.

Baselines. To verify the performance of our framework, we compare it against a comprehensive set of baselines categorized into two groups:

(1) Standard Multimodal Learning, under complete modalities, including TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), MMIM (Han et al., 2021), EMOE (Fang et al., 2025) and MMA (Chen et al., 2025).

(2) Imperfect Multimodal Learning, which is specifically designed for missing or noisy scenarios, including MulT (Tsai et al., 2019), MCTN (Pham et al., 2019), MISA (Hazarika et al., 2020), MMIN (Zhao et al., 2021), C-MIB (Mai et al., 2023), IMDER (Wang et al., 2023), Multimodal-Boosting (Mai et al., 2024), MoMKE (Xu et al., 2024), SAM-LML (Mai et al., 2025) and PaSE (He et al., 2025a).

Evaluation Metrics. For CMU-MOSI and CMU-MOSEI, we follow (Tsai et al., 2019) to evaluate our method by using the metrics: 7-class Accuracy (ACC_7), Binary Accuracy (ACC_2), F1-score (F_1), and Mean-absolute Error (MAE). For MIntRec, we follow the standard protocol (Sun et al., 2024) to evaluate the results via: ACC and F_1. For IEMOCAP (Liang et al., 2021), we use the average ACC and F_1 as evaluation metrics.

Implementation Details. All models are implemented using PyTorch and trained on six NVIDIA RTX 4090 GPUs. We employ the Adam optimizer with a dropout rate of 0.1 to prevent overfitting. For QA-MoE, we construct the expert bank with $N = 8$ GLU-based experts. The dual-path router is configured to activate the top-k ($k = 3$) experts for sparse computation.

4.2 Performance on Standard Benchmarks

To ensure a fair evaluation of the architectural effectiveness, both the baselines and QA-MoE are trained on the original clean datasets without any degradation injection. Table 1 presents the performance comparison on aligned CMU-MOSI, CMU-MOSEI(unaligned is shown in Appendix A.4.1) and MIntRec. On CMU-MOSI, our method surpasses the MMA by a improvement of **6.7%** in ACC_7 and **1.2%** in F_1 score. On other datasets, QA-MoE also outperforms prior methods consistently. It also indicates that the advantages are from the intrinsic design of our framework rather than data augmentation strategies.

Models	CMU-MOSI					CMU-MOSEI					MIntRec	
	$ACC_7 \uparrow$	$ACC_2 \uparrow$	$F_1 \uparrow$	MAE $\downarrow$	Corr $\uparrow$	$ACC_7 \uparrow$	$ACC_2 \uparrow$	$F_1 \uparrow$	MAE $\downarrow$	Corr $\uparrow$	ACC $\uparrow$	$F_1 \uparrow$
TFN†	31.9	78.8	78.9	0.953	0.698	50.9	80.4	80.7	0.574	0.700	-	-
LMF†	36.9	78.7	78.7	0.931	0.695	52.3	84.7	84.5	0.564	0.677	-	-
MulT†	35.1	80.0	80.1	0.936	0.711	52.3	82.7	82.8	0.572	-	72.6	69.5
MISA†	41.8	84.2	84.2	0.754	0.761	52.3	85.3	85.1	0.543	0.756	72.4	70.8
MMIM†	45.8	84.6	84.5	0.717	-	50.1	83.6	83.5	0.580	-	-	-
EMOE†	47.7	85.4	85.4	0.710	-	54.1	85.3	85.3	0.536	-	72.6	70.7
MMA‡	46.9	86.4	86.4	0.693	0.803	55.2	85.7	85.7	0.529	0.766	-	-
Ours	**53.6**	**88.2**	**87.6**	**0.579**	**0.817**	**58.4**	**87.1**	**87.1**	**0.477**	**0.791**	**75.3**	**72.2**

Table 1: Experimental results on CMU-MOSI, CMU-MOSEI, and MIntRec datasets. The results marked with † are retrieved from Fang et al. (2025), and those with ‡ are cited from Chen et al. (2025).

Datasets	Models	Testing Condition (Available Modalities)							Avg.
		$\{t\}$	$\{a\}$	$\{v\}$	$\{t,a\}$	$\{t,v\}$	$\{a,v\}$	$\{t,a,v\}$	
IEMOCAP	MulT	62.4 / 63.7	49.7 / 51.6	48.9 / 45.7	68.3 / 69.4	67.8 / 68.3	56.3 / 55.8	70.1 / 70.5	60.5 / 60.7
	MISA	66.5 / 68.0	56.5 / 59.0	52.5 / 51.6	72.9 / 75.1	72.6 / 73.6	63.9 / 65.4	74.2 / 74.5	65.5 / 66.7
	MMIM	67.0 / 68.2	55.0 / 53.2	51.9 / 50.4	74.0 / 75.4	72.6 / 73.6	65.3 / 66.5	75.5 / 75.8	65.9 / 66.1
	Ours	**71.2 / 72.3**	**58.2 / 59.1**	**54.6 / 53.8**	**75.1 / 75.3**	**73.8 / 74.1**	**66.9 / 66.8**	**77.1 / 77.3**	**68.1 / 68.4**
CMU-MOSI	MCTN	79.10 / 79.20	56.10 / 54.50	55.00 / 54.40	81.00 / 81.00	81.10 / 81.20	57.50 / 57.40	81.40 / 81.50	68.30 / 67.95
	MMIN	83.80 / 83.80	55.30 / 51.50	57.00 / 54.00	84.00 / 84.00	83.80 / 83.90	60.40 / 58.50	84.60 / 84.40	72.72 / 69.28
	IMDer	84.80 / 84.70	62.00 / 62.20	61.30 / 60.80	85.40 / 85.30	85.50 / 85.40	63.60 / 63.40	85.70 / 85.60	73.77 / 73.63
	MoMKE	86.59 / 86.52	63.19 / 58.61	**63.35 / 63.34**	87.20 / 87.17	87.04 / 87.00	64.04 / 64.66	87.96 / 87.89	75.24 / 74.55
	PaSE	84.70 / 84.23	60.01 / 58.79	61.43 / 61.50	86.71 / 86.79	87.14 / 86.99	63.35 / 63.32	88.32 / 88.25	73.89 / 73.60
	Ours	**87.24 / 87.35**	**63.73 / 60.71**	62.58 / 62.42	**88.51 / 87.64**	**88.11 / 88.15**	**65.69 / 65.20**	**89.97 / 89.02**	**77.98 / 77.21**
CMU-MOSEI	MCTN	82.60 / 82.80	62.70 / 54.50	62.60 / 57.10	83.50 / 83.30	83.20 / 83.20	63.70 / 62.70	84.20 / 84.20	73.05 / 70.60
	MMIN	82.30 / 82.40	58.90 / 59.50	59.30 / 60.00	83.70 / 83.30	83.80 / 83.40	63.50 / 61.90	84.30 / 84.20	71.92 / 71.75
	IMDer	84.50 / 84.50	63.80 / 60.60	63.90 / 63.60	85.10 / 85.10	85.00 / 85.00	64.90 / 63.50	85.10 / 85.10	76.00 / 75.30
	MoMKE	86.46 / 86.43	72.56 / 71.03	70.12 / 70.23	86.68 / 86.61	86.79 / 86.69	**73.34 / 71.82**	87.12 / 87.03	79.33 / 78.80
	PaSE	84.36 / 84.08	69.04 / 68.56	68.69 / 68.74	86.47 / 86.42	86.73 / 86.45	72.03 / 71.90	88.10 / 87.96	77.89 / 77.69
	Ours	**87.61 / 87.57**	**73.01 / 72.77**	**71.38 / 71.07**	**87.78 / 87.78**	**87.91 / 87.89**	73.24 / 73.19	**88.93 / 89.01**	**81.41 / 81.33**

Table 2: Performance comparison under various modality missingness scenarios. The values denote ACC_2/F_1. The testing conditions indicate the **available** modalities (e.g., $\{t\}$ means only Textual is available).

4.3 Performance under Specific Imperfections

4.3.1 Evaluation on Protocol I

To evaluate the model robustness against modality missingness, we adopt **Protocol I**. ACCording to (Wang et al., 2023), it has been divided into two commonly-used protocols.:

Fixed Missing Protocol. It is designed to stimulate permanent sensor failure during inference. We test the models on all possible subsets of modalities obtained from the original modality dataset. The results in Table 2 reveal that baseline models suffer great degradation without *text*. In contrast, QA-MoE exhibits remarkable stability. Taking the $\{a, v\}$ setting as an example, the quality-aware router implicitly detects the missing text modality as having extreme aleatoric uncertainty. Consequently, the quality score r_t asymptotically approaches zero, which automatically reduces the uninformative text branch. It proves that our model does not rely on a single modality but treats all modalities independently.

Random Missing Protocol. Consistent with (Lian et al., 2023), we keep the same missing rate η during training, validation, and testing phases. Table 3 reports the performance under varying missing rates η from 10% to 70%. Compared with SAM-LML dropping 19.2% on CMU-MOSI as η shifts from 10% to 70%, QA-MoE achieves an average ACC_7 of **42.0%**, which surpasses the strongest baseline by a substantial margin of **6.1%**. It indicates that for each specific sample, our model can focus on the existing modalities, ensuring the high-fidelity inference in severe data sparsity.

4.3.2 Evaluation on Protocol II

To comprehensively evaluate model robustness against modality noise, we adopt the diverse noise injection protocol established in (Mai et al., 2025). The Noise Intensity λ varies from 0.1 to 0.7, controlling the intensity of the degradation. We compare QA-MoE against state-of-the-art robust frame-

Dataset	Models	Random Missing Rate (η)							Avg.
		10%	20%	30%	40%	50%	60%	70%	
CMU-MOSI	MCTN	39.8 / 78.5	38.5 / 75.7	35.5 / 71.2	32.9 / 67.6	31.2 / 64.8	29.7 / 62.5	27.5 / 59.0	33.6 / 68.5
	MMIN	41.2 / 81.8	38.9 / 79.1	36.9 / 76.2	34.9 / 71.6	34.2 / 66.5	29.1 / 64.0	28.4 / 61.0	34.5 / 71.5
	IMDer	42.1 / 83.4	41.6 / 80.5	37.4 / 77.6	35.2 / 66.3	29.5 / 65.4	27.0 / 65.5	26.5 / 60.4	34.2 / 71.3
	MoMKE	35.1 / 81.6	32.9 / 76.6	30.6 / 71.7	28.4 / 67.5	26.2 / 63.2	23.9 / 58.9	22.4 / 55.9	28.5 / 67.9
	SAM-LML	45.6 / 84.7	42.9 / 81.2	37.5 / 78.1	37.8 / 74.7	32.8 / 70.9	28.1 / 66.6	26.4 / 65.6	35.9 / 74.5
	Ours	**53.3 / 85.1**	**51.2 / 81.5**	**47.2 / 78.4**	**40.5 / 76.4**	**37.2 / 73.2**	**33.9 / 69.1**	**30.5 / 68.7**	**42.0 / 76.1**
CMU-MOSEI	MCTN	49.8 / 81.6	48.6 / 78.7	47.4 / 76.2	45.6 / 74.1	45.1 / 72.6	43.8 / 71.1	43.6 / 70.5	46.3 / 75.0
	MMIN	50.6 / 81.3	49.6 / 78.8	48.1 / 75.5	47.5 / 72.6	46.7 / 70.7	45.6 / 70.3	44.8 / 69.5	47.6 / 74.1
	IMDer	52.1 / 82.9	51.3 / 79.7	49.6 / 77.8	48.0 / 73.3	46.6 / 68.4	45.0 / 65.9	44.1 / 66.6	48.1 / 73.5
	MoMKE	47.2 / 84.7	45.4 / 82.7	43.6 / 80.7	41.7 / 78.7	39.8 / 76.7	37.9 / 74.7	36.7 / 73.3	41.8 / 78.8
	SAM-LML	51.9 / 84.5	51.7 / 83.7	48.7 / 81.6	48.3 / 79.5	46.9 / 77.4	45.6 / 76.4	44.5 / 74.6	48.2 / 79.7
	Ours	**55.8 / 86.4**	**54.2 / 81.9**	**51.2 / 80.3**	**50.3 / 79.9**	**48.3 / 78.7**	**47.4 / 77.0**	**46.1 / 75.3**	**50.5 / 79.9**

Table 3: Robustness comparison under **Random Missing Protocol** ($\mathrm{ACC}_7/\mathrm{F}_1$).

Dataset	NR	C-MIB ACC_2 / MAE	MM-Boosting ACC_2/ MAE	SAM-LML ACC_2 / MAE	QA-MoE (Ours) **ACC_2 / MAE**
CMU-MOSI	0.1	87.8 / 0.670	86.7 / 0.678	88.4 / 0.636	**89.4 / 0.616**
	0.2	87.5 / 0.726	86.1 / 0.738	88.1 / 0.665	**88.9 / 0.636**
	0.3	86.4 / 0.912	86.4 / 0.785	87.8 / 0.663	**88.6 / 0.639**
	0.4	83.2 / 1.366	85.5 / 0.841	87.6 / 0.666	**88.4 / 0.641**
	0.5	84.9 / 1.660	86.1 / 1.172	88.1 / 0.666	**88.1 / 0.649**
	0.6	80.8 / 2.595	82.0 / 1.355	87.5 / 0.660	**87.8 / 0.652**
	0.7	82.1 / 3.146	84.4 / 1.750	87.3 / 0.669	**87.7 / 0.660**
	Avg.	84.7 / 1.582	85.3 / 1.046	87.8 / 0.661	**88.4 / 0.642**
CMU-MOSEI	0.1	86.1 / 0.545	86.4 / 0.544	87.0 / 0.521	**88.2 / 0.498**
	0.2	84.5 / 0.582	86.6 / 0.557	87.0 / 0.525	**88.1 / 0.501**
	0.3	85.6 / 0.622	85.5 / 0.623	87.3 / 0.522	**88.0 / 0.512**
	0.4	84.4 / 0.703	85.3 / 0.682	87.2 / 0.525	**87.6 / 0.514**
	0.5	83.7 / 0.875	84.1 / 0.724	86.6 / 0.529	**87.4 / 0.521**
	0.6	82.4 / 1.054	85.4 / 0.924	87.0 / 0.532	**87.2 / 0.523**
	0.7	80.5 / 1.404	80.3 / 1.125	85.9 / 0.545	**86.8 / 0.531**
	Avg.	83.9 / 0.826	84.8 / 0.740	86.9 / 0.528	**87.6 / 0.514**

Table 4: Comparison under varying NI (λ).The results are cited from (Mai et al., 2025).

Model	Training Strategy	Mixed Test Set (Protocol III) MAE ↓	ACC_7 ↑	ACC_2 ↑	F_1 ↑
MMA	Clean Train	0.693	46.9	86.4	86.4
	Spectrum Train	0.688	51.8	87.2	88.4
SAM-LML	Clean Train	0.628	49.4	89.2	89.1
	Spectrum Train	0.599	52.5	89.6	89.7
QA-MoE (Ours)	Clean Train	0.589	53.6	87.3	87.6
	Spectrum Train	**0.515**	**54.5**	**87.9**	**89.1**

Table 5: Decoupling analysis under Protocol III.

works using their reported settings. Table 4 visualizes the performance trends on CMU-MOSI and CMU-MOSEI. As the noise intensity increases, standard baselines exhibit rapid performance decay. While SAM-LML shows improved resistance, QA-MoE consistently outperforms all baselines across the entire noise spectrum. Notably, at severe noise levels ($\lambda = 0.7$), QA-MoE maintains a lead of roughly **0.9%** over SAM-LML, which validates that our distributional reliability scoring effectively filters out high-variance features and reconstructs semantics from noisy signals.

4.4 Performance on Spectrum Dataset

Moving beyond standard benchmarks, we evaluate the **Protocol III** under the proposed Continuous Reliability Spectrum. The model encounters a heterogeneous test set comprising random combinations of noise (λ) and missingness (η), simulating the unpredictable imperfections of real-world deployment.

4.4.1 Analysis of Architectural Superiority

A fundamental question arises regarding the source of our model's efficacy: *Does the robustness of QA-MoE stem from its intrinsic architectural design, or merely from the Spectrum-Aware training strategy?* To rigorously disentangle these factors, we conduct a controlled experiment by retraining the strongest baselines using the exact same dynamic degradation injection strategy employed in our framework. Table 5 reports the performance on the heterogeneous Protocol III test set. Remarkably, the QA-MoE trained solely on clean data (53.6%) still outperforms the strongest baseline enhanced by the spectrum training strategy (52.5%). It convinces that our robustness derives primarily from the proposed mechanism in handling unseen shifts, rather than relying on data augmentation.

4.4.2 Reliability Score Validation

To verify that the learned quality score r_m faithfully tracks real-world signal degradation, we conduct two quantitative analyses on CMU-MOSI.

Correlation with Ground-Truth Corruption. We compute the Spearman rank correlation ρ be-

r_m	[0.0, 0.2)	[0.2, 0.4)	[0.4, 0.6)	[0.6, 0.8)	[0.8, 1.0]
Avg.	1.42	1.15	0.92	0.74	0.58

Table 6: Alignment between reliability score r_m and prediction error (MAE) on CMU-MOSI. Higher r_m consistently corresponds to lower prediction error, validating r_m as an effective quality indicator.

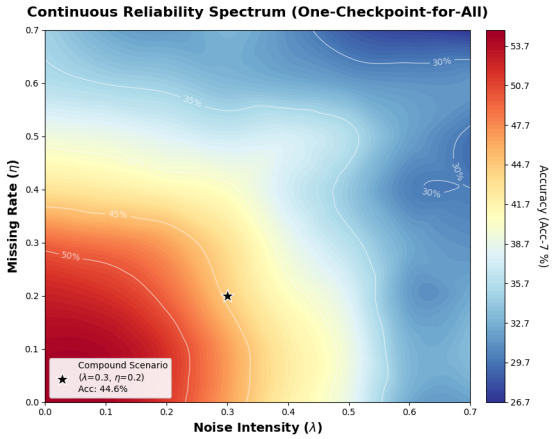

Figure 3: **Continuous Reliability Landscape.** The smooth performance gradient (from warm to cool colors) demonstrates that QA-MoE exhibits graceful degradation rather than abrupt failure. The star ($\star$) marks the compound defect scenario ($\lambda = 0.3, \eta = 0.2$).

tween r_m and the ground-truth degradation levels (λ for noise intensity, η for missing rate). The results yield $\rho = -0.87$ ($p < 0.001$), indicating a strong and statistically significant negative correlation: as input quality deteriorates, r_m monotonically decreases. This confirms that the self-supervised heteroscedastic objective in Eq. **??** drives σ^2— and consequently r_m—to serve as a reliable proxy for input-dependent signal quality, without requiring any explicit quality labels.

Alignment with Prediction Error. Beyond correlation, we further validate whether r_m reflects actual predictive risk. We partition test samples into five bins by r_m and report the average MAE within each bin (Table 6). As shown, prediction error decreases monotonically as r_m increases, dropping from 1.42 (low reliability) to 0.58 (high reliability). This alignment confirms that r_m is a high-quality indicator of modality reliability, effectively guiding the router to suppress noisy inputs.

4.4.3 "One-Checkpoint-for-All" Strategy

Unlike existing methods often require retraining to adapt to specific noise levels, QA-MoE is designed to remain effective across varying reliability conditions. To verify this, we evaluate a single trained checkpoint across the entire reliability spectrum grid without any parameter tuning. Figure 3 visualizes the model's performance on the Continuous Reliability Spectrum introduced in Figure 1. Unlike discrete evaluations, this continuous surface demonstrates a high-fidelity plateau covering the Stochastic Mixture region. The model maintains robustness well into the degradation zones ($\star$), confirming its ability to adapt continuously to unseen defects. The detailed value of each discrete points are also provided in Appendix A.4.2. Besides, it also provides the performance under other SAM-LML, which shows that our model can maintain effective across the spectrum without retraining.

4.5 Model Analysis

In this section, we conduct a comprehensive analysis to provide deeper insights into the properties of QA-MoE. Besides the following, the analysis of the computational efficiency is placed in Appendix A.5 due to space.

Model Variants	**Spectrum Training Setting**			
	MAE ↓	ACC_7 ↑	ACC_2 ↑	F_1 ↑
QA-MoE (Full)	**0.525**	**54.5**	**88.8**	**89.1**
w/o Quality Gating	0.884	52.1	76.2	77.4
w/o Variance ($\boldsymbol{\sigma}^2$)	0.795	50.7	79.1	51.2
w/o Universal Fallback ($\mathbf{y}_{prior}$)	0.780	48.6	78.5	45.9

Table 7: Ablation studies on CMU-MOSI. We evaluate the contribution of the key parts of our QA-MoE.

Ablation Study. To disentangle component contributions, we conduct ablation studies on CMU-MOSI (Table 7) with Spectrum Training. First, removing the quality gate significantly increases errors on noisy inputs, confirming the necessity of explicit signals to bypass unreliable experts. Second, relying solely on the mean vector causes performance degradation, validating that the second moment is a critical proxy for aleatoric uncertainty.
Parameter Sensitivity Analysis. We conduct sensitivity analysis on the number of active experts (k). We fix the total number of experts $N = 8$ and vary the active selection $k \in \{1, 2, 3, 4, 8\}$. Figure 4 shows a performance peak at $k = 3$. The tendency indicates that a single expert is insufficient to capture complex multimodal dynamics and increasing k to 8 results in degradation due to overfitting. Thus, $k = 3$ represents the optimal trade-off between effectiveness and efficiency.
Interpretability Analysis. Figure 5 shows the reconfiguration of expert attention under varying

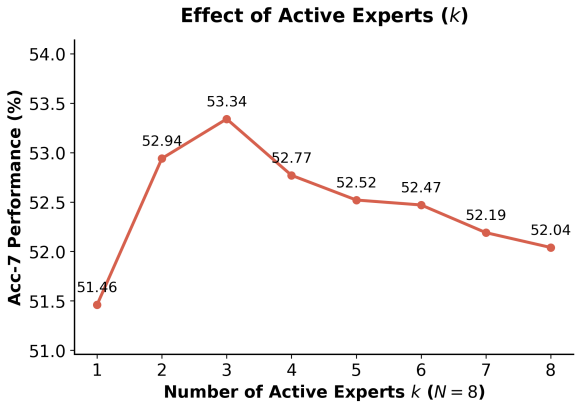

Figure 4: Parameter k Sensitivity Analysis.

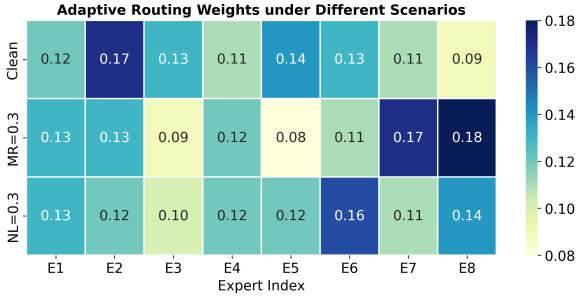

Figure 5: Visualization of Adaptive Routing Patterns.

degradation contexts. We observe a distinct quality-aware routing shift. Notably, Expert 2 dominates in clean settings but is suppressed under corruption to prevent error propagation. In contrast, Expert 8 gains prominence specifically under imperfect scenarios. Besides, the divergent behavior in Expert 5 under different conditions confirms that QA-MoE discriminates between specific failure modes rather than applying a generic penalty.

Discussion on Universal Fallback Mechanism. While the ablation in Table 7 confirms the necessity of y_{prior}, we further analyze the design trade-off it introduces. The interpolation is intentional: under severe corruption ($r_m \to 0$), the model deliberately falls back to a dataset-level semantic consensus rather than propagating unreliable expert outputs. However, this design implies that y_{prior} must capture a meaningful semantic prior. To verify this, we examine the learned y_{prior} across different dataset splits and find it remains consistent, suggesting it encodes a stable global representation rather than overfitting to training distribution. The performance drop observed when removing y_{prior} (Table 7) therefore reflects the importance of this safety net, particularly under high degradation rates where instance-specific signals become unreliable.

5 Conclusion

In this work, we introduce the **Continuous Reliability Spectrum** to model real-world imperfections and propose **QA-MoE** to address them. By dynamically routing signals based on self-supervised aleatoric uncertainty, QA-MoE achieves a *One-Checkpoint-for-All* capability across diverse degradation protocols, establishing state-of-the-art performance on multiple benchmarks. A remaining risk is that the routing effectiveness relies on the precision of quality predictors, where estimation errors in extreme edge cases could lead to suboptimal performance. Beyond the current scope, we note that the probabilistic feature modeling and quality-aware routing in QA-MoE are architecture-agnostic: the reliability score r_m derived from per-modality aleatoric uncertainty can be applied to any high-dimensional embeddings, including those from large multimodal models. This suggests a promising direction of using QA-MoE as a plug-and-play robust adapter within LMM pipelines to enhance their resilience to noisy or incomplete real-world inputs. To support reproducibility and future research, we will release the complete source code and model checkpoints upon acceptance.

6 Acknowledgement

The work is mainly supported by Education Bureau of Guangzhou Municipality No. 2024312045. The work also receives support from the AI Research and Learning Base of Urban Culture under Project 2023WZJD008.

Limitations

Despite the effectiveness of our approach, there are two main limitations. First, the quality signals in our framework are learned implicitly in a self-supervised manner; thus, the model lacks explicit interpretability regarding specific noise types (e.g., blur vs. occlusion). Second, the MoE architecture involves a routing mechanism that introduces additional computational overhead compared to static fusion methods. We plan to explore more efficient routing strategies and fine-grained quality modeling in future work. Finally, while y_{prior} serves as an effective universal fallback, the model's reliance on this learned static embedding may limit its adaptability when the test distribution diverges significantly from the training set in cross-domain or few-shot scenarios.

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

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

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1158–1166, New York, NY, USA. Association for Computing Machinery.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

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

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736, Bangkok, Thailand. Association for Computational Linguistics.

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

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Kang He, Boyu Chen, Yuzhe Ding, Fei Li, Chong Teng, and Donghong Ji. 2025a. Pase: Prototype-aligned calibration and shapley-based equilibrium for multimodal sentiment analysis. *Preprint*, arXiv:2511.17585.

Xilin He, Haijian Liang, Boyi Peng, Weicheng Xie, Muhammad Haris Khan, Siyang Song, and Zitong Yu. 2025b. Msamba: Exploring multimodal sentiment analysis with state space models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1309–1317.

Guanxuan Jiang, Shirao Yang, Yuyang Wang, and Pan Hui. 2026. When trust collides: Exploring human-llm cooperation intention through the prisoner's dilemma. *International Journal of Human-Computer Studies*, page 103740.

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

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, and 1 others. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. *Advances in neural information processing systems*, 2021(DB1):1.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Mengmeng Ma, Jian Ren, Long Zhao, S. Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. *ArXiv*, abs/2103.05677.

Sijie Mai, Shiqin Han, and Haifeng Hu. 2025. Supervised attention mechanism for low-quality multimodal data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21377–21397, Suzhou, China. Association for Computational Linguistics.

Sijie Mai, Ya Sun, Aolin Xiong, Ying Zeng, and Haifeng Hu. 2024. Multimodal boosting: Addressing noisy modalities and identifying modality contribution. *IEEE Transactions on Multimedia*, 26:3018–3033.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2023. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *Trans. Multi.*, 25:4121–4134.

David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 2023. 4m: massively multimodal masked modeling. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: learning robust joint representations by cyclic translations between modalities. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Chengkai Shi and Yunhua Zhang. 2025. Mmkt: Multimodal sentiment analysis model based on knowledge-enhanced and text-guided learning. *Applied Sciences*, 15(17).

Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. 2024. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26963–26973.

Kaiwei Sun and Mi Tian. 2025. Sequential fusion of text-close and text-far representations for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 40–49, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023. Incomplete multimodality-diffused emotion recognition. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. 2024. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 438–446, New York, NY, USA. Association for Computing Machinery.

Yang Yang, Xunde Dong, and Yupeng Qiang. 2024. CLGSI: A multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2099–2110, Mexico City, Mexico. Association for Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, E. Cambria, and Louis philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1688–1697, New York, NY, USA. Association for Computing Machinery.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Haoyu Zhang, Yinan Zhang, Chaolong Ying, Xiaoying Tang, and Tianshu Yu. 2025. Improving task-specific multimodal sentiment analysis with general mllms via prompting. In *Advances in Neural Information Processing Systems*. NeurIPS 2025.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. 2023. Meta-transformer: A unified framework for multimodal learning. *Preprint*, arXiv:2307.10802.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, Online. Association for Computational Linguistics.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025a. Multimodal invariant sentiment representation learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14743–14755, Vienna, Austria. Association for Computational Linguistics.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *Preprint*, arXiv:2310.01852.

Yitong Zhu, Lei Han, Guanxuan Jiang, PengYuan Zhou, and Yuyang Wang. 2025b. Hierarchical moe: Continuous multimodal emotion recognition with incomplete and asynchronous inputs. *Preprint*, arXiv:2508.02133.

A Appendix

A.1 Spectrum Dataset Generation Details

To implement the **Spectrum-Aware Training Strategy** described in Section 3.5, we construct a dynamic data loader that applies stochastic transformations on-the-fly. Unlike static augmentation, this process generates a unique view of the dataset for every training batch, ensuring the model traverses the continuous reliability spectrum.

A.1.1 Dynamic Injection Protocol

To ensure the model generalizes across the entire reliability spectrum, we employ a batch-wise dynamic sampling strategy. Specifically, at the beginning of each training iteration, we sample a pair of global degradation coefficients $(\lambda_{batch}, \eta_{batch})$ from a uniform distribution:

$$\lambda_{batch} \sim \mathcal{U}(0,1), \quad \eta_{batch} \sim \mathcal{U}(0,1) \tag{11}$$

These coefficients are then applied uniformly to all samples within the current mini-batch. This exposes the router to continuously varying difficulty levels throughout the training epochs, preventing overfitting to any specific discrete noise intensity.

A.1.2 Modality-Specific Degradation

Since the modalities (Text, Audio, Vision) have distinct physical properties, we design specific degradation functions $\mathcal{T}(\cdot)$ for each, consistent with the Stochastic Imperfection Modeling in Eq. 2.

Continuous Modalities (Audio & Vision). For the continuous feature vectors from acoustic (e.g., COVAREP/Wav2Vec) and visual (e.g., Facet/ViT) encoders, we verify robustness by injecting additive noise. The corrupted feature $\tilde{\mathbf{u}}_m$ is generated as:

$$\tilde{\mathbf{u}}_m = \mathbf{u}_m + \boldsymbol{\epsilon}_m, \quad \boldsymbol{\epsilon}_m \sim \mathcal{N}(\mathbf{0}, (\lambda \cdot \sigma_{\text{ref}})^2 \mathbf{I}) \tag{12}$$

where σ_{ref} is a reference standard deviation calculated from the training set statistics to ensure the noise scale is relative to the feature.

Acoustic: We simulate background noise and sensor jitter using Additive White Gaussian Noise (AWGN).

Visual: We simulate blur and low-light sensor noise. While actual blur is a convolution operation, in the high-level feature space, this is effectively modeled by increasing the feature variance via additive Gaussian noise.

Discrete Modality (Text). For textual data, noise manifests as Automatic Speech Recognition (ASR) errors or missing words. We implement this via a **Token-Level Dropout** mechanism. Given a sequence of word embeddings $\mathbf{u}_t = \{w_1, w_2, \ldots, w_L\}$, each token is independently replaced by a zero vector (or a special `[MASK]` token) with probability $p = \lambda$:

$$\tilde{w}_i = \begin{cases} w_i & \text{with probability } 1-\lambda \\ \mathbf{0} & \text{with probability } \lambda \end{cases} \tag{13}$$

This simulates semantic fragmentation ranging from minor typos (low λ) to unreadable sentences (high λ).

Modality Missingness. Finally, to simulate complete sensor failure (Protocol I), we apply the global missingness mask $\mathbb{I}_{\text{miss}}$. With probability η, the entire feature sequence for a modality m is zeroed out: $\tilde{\mathbf{u}}_m \leftarrow \mathbf{0}$.

A.2 Datasets and Feature Extraction

CMU-MOSI (Zadeh et al., 2016) and **CMU-MOSEI** (Bagher Zadeh et al., 2018) are the most widely used benchmarks for MSA and MER tasks. CMU-MOSI consists of 2,199 opinion video clips labeled with sentiment intensity scores ranging from -3 (highly negative) to +3 (highly positive). CMU-MOSEI is a larger-scale dataset containing 23,453 annotated video segments. Both datasets are pre-processed and word-aligned following the standard protocol. We strictly follow the standard feature extraction protocols established in prior literature (Tsai et al., 2019; Hazarika et al., 2020), and we utilize 300-dimensional GloVe language features (Pennington et al., 2014) and 768-dimensional BERT-base-uncased hidden states (Devlin et al., 2019). Facet (Baltrušaitis et al., 2016) provides 35 facial action unit visual features, and COVAREP (Degottex et al., 2014) offers 74-dimensional acoustic features.

IEMOCAP (Busso et al., 2008) is a multimodal database for emotion recognition, comprising dyadic conversations between ten speakers. Following prior works (Tsai et al., 2019), we focus on the classification of six discrete emotions: happy, sad, angry, fearful, frustrated, and neutral. For IEMOCAP, we follow (Zhao et al., 2021) to extract acoustic, visual and textual features.

MIntRec (Zhang et al., 2022) is a challenging dataset for multimodal intent recognition capturing

Models	CMU-MOSI					CMU-MOSEI				
	ACC-7 ↑	ACC-2 ↑	F1 ↑	MAE ↓	Corr ↑	ACC-7 ↑	ACC-2 ↑	F1 ↑	MAE ↓	Corr ↑
TFN†	35.3	76.5	76.6	0.995	0.698	50.2	84.2	84.0	0.573	0.700
LMF†	31.1	79.1	79.1	0.963	0.695	51.9	83.8	83.9	0.565	0.677
MulT†	33.2	80.3	80.3	0.933	0.711	53.2	84.0	84.0	0.556	-
MISA†	43.6	83.8	83.9	0.742	0.761	51.0	84.8	84.8	0.557	0.756
MMIM†	45.9	83.4	83.4	0.777	-	52.6	81.5	81.3	0.578	-
EMOE†	47.8	85.4	85.3	0.697	-	53.9	85.5	85.5	0.530	-
Ours	**53.3**	**87.4**	**87.2**	**0.583**	**0.816**	**58.4**	**87.1**	**87.1**	**0.477**	**0.791**

Table 8: Experimental results on **CMU-MOSI** and **CMU-MOSEI** datasets. The results marked with † are retrieved from Fang et al. (2025), and those with ‡ are cited from Chen et al. (2025).

MR (η)	**Noise Intensity** (λ)							
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0%	**54.50**	54.27	51.31	46.65	41.98	38.63	32.36	31.92
10%	54.20	53.98	51.60	46.79	42.27	39.07	32.94	33.38
20%	52.60	51.79	49.27	**44.61**	40.67	37.76	31.63	32.36
30%	49.25	48.31	46.65	43.15	38.92	35.71	32.22	31.05
40%	43.06	42.38	41.44	40.38	37.17	34.11	30.32	30.03
50%	39.31	38.85	37.90	36.94	37.32	35.57	31.92	29.59
60%	35.69	34.94	34.40	33.23	32.59	31.71	31.20	31.34
70%	34.69	32.34	31.55	32.80	31.03	28.53	27.47	26.88

Table 9: **Discrete Evaluation Grid (ACC-7 %).** This table presents the exact performance metrics of the single QA-MoE checkpoint across varying degrees of degradation. The gray cell marks the compound defect scenario ($\lambda = 0.3, \eta = 20\%$) analyzed in Figure 3.

high-quality "in-the-wild" interactions. Unlike lab-controlled datasets, MIntRec naturally contains environmental noise and diverse background scenes, making it an ideal testbed for evaluating model robustness against real-world imperfections. On MIntRec, dimensions for text, visual, and acoustic features are 768, 256, and 768, respectively.

A.3 Baslines, Implementation and Metrics

Baselines. To verify the effectiveness of our framework, we compare it against a comprehensive set of baselines categorized into two groups:

General Multimodal Learning Approaches: We select methods that focus on sophisticated fusion mechanisms assuming complete modalities. These include TFN (Zadeh et al., 2017) and LMF (Liu et al., 2018) which utilize tensor fusion; MulT (Tsai et al., 2019) which employs cross-modal transformers; MISA (Hazarika et al., 2020) which focuses on feature disentanglement; MMIM (Han et al., 2021) which maximizes mutual information; and inspired by Mixture of Experts, EMOE (Fang et al., 2025) and MMA (Chen et al., 2025) facilitate adaptive multimodal fusion.

Robustness-Oriented Approaches: We also compare against methods specifically designed for handling missing or noisy modalities. These include MMIN (Zhao et al., 2021)which reconstructs missing modalities via cascaded prediction, and SMIL (Ma et al., 2021) which utilizes Bayesian meta-learning to handle severe modality absence.

Implementation. We implement all models using PyTorch on NVIDIA RTX 4090 GPUs. Following standard protocols (Tsai et al., 2019), we utilize BERT-base-uncased ($d_t = 768$) for text, and acoustic/visual features extracted via COVAREP ($d_a = 74$) and Facet ($d_v = 35$), respectively. The hidden dimension of the multimodal encoders is set to $d_{model} = 128$. Models are trained for 30 epochs with a batch size of 16. We employ the Adam optimizer with $\beta = (0.9, 0.999)$ and a weight decay of $1e^{-5}$. To prevent overfitting, we apply a dropout rate of 0.1 and gradient clipping (threshold 1.0). The learning rate is tuned via grid search within $\{1e^{-3}, 5e^{-4}, 1e^{-4}\}$ and decayed using a Cosine Annealing scheduler. For the QA-MoE structure, we set the total experts $N = 8$ and active experts $k = 3$. The Quality Predictors are implemented as two-layer MLPs to ensure lightweight computation. For strict reproducibility, all experiments are conducted with a fixed random seed (1111).

A.4 Supplementary Experimental Results

A.4.1 Results on Perfect Dataset

We evaluate the model under the challenging unaligned setting, where modalities possess inherent temporal asynchrony. Compared to strong baselines, including the recent MoE-based method EMOE (Fang et al., 2025), QA-MoE achieves substantial gains across all metrics. On **CMU-MOSI**, we achieve an ACC_7 of **53.3%**, surpassing the previous best (EMOE) by **+5.5%**. On **CMU-MOSEI**, we get an ACC_7 of **58.4%**, outperforming the strongest baseline by **+4.5%**. These results confirm that our Quality-Aware Routing mechanism is not solely a defensive measure against noise. Even in perfect datasets, the router effectively dynamically selects experts to handle the natural semantic misalignment and heterogeneity inherent in unaligned multimodal settings which proves the architecture's intrinsic superiority.

A.4.2 Results on Spectrum Dataset

To ensure reproducibility and transparency, we provide the comprehensive numerical results corresponding to the One-Checkpoint-for-All evaluation discussed in Section 4.3.2. Table 9 details the ACC_7 performance of QA-MoE across the complete discrete grid of Noise Intensities ($\lambda \in [0, 0.7]$) and Missing Rates ($\eta \in [0, 70\%]$). These raw values serve as the basis for the continuous landscape visualization in Figure 3. Notably, the table confirms the model's graceful degradation:

(1) Under **Ideal Conditions** ($\lambda = 0, \eta = 0$), the model achieves a peak accuracy of **54.50%**.

(2) Under the **Compound Scenario** highlighted in the main text ($\lambda = 0.3, \eta = 20\%$), the model retains a robust accuracy of **44.61%**, validating the effectiveness of the quality-aware routing mechanism even when subject to simultaneous mixed imperfections.

To contrast with the stability of QA-MoE, we visualize the reliability landscape of the strongest baseline, SAM-LML, in Figure 6. Unlike our proposed method, which maintains a high-performance plateau, SAM-LML exhibits significant **brittleness** to unseen degradation.As observed, the high-accuracy region (red/orange) is strictly confined to the top-left corner (clean data). A minor shift into the mixed degradation zone triggers a drastic performance drop-off. For instance, at the marked compound defect point ($\lambda = 0.3, \eta = 0.2$), the accuracy collapses to **35.5%**,

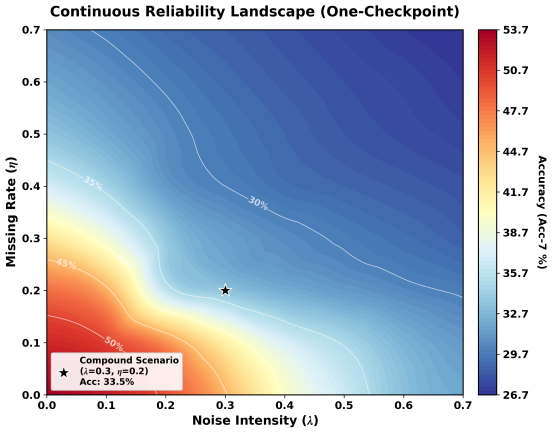

Figure 6: **Reliability Landscape of Baseline (SAM-LML).** The visualization reveals a sharp performance decay, forming a reliability cliff." While the model achieves peak performance at the clean origin, its accuracy plummets rapidly as degradation intensity increases. The star ($\star$) marks the compound defect scenario ($\lambda = 0.3, \eta = 0.2$), where accuracy has already degraded to 35.5%, demonstrating the lack of robustness in the One-Checkpoint" setting.

representing a loss of over **15%** compared to its clean performance. This confirms that without the dynamic expert routing mechanism, conventional models cannot effectively support the "One-Checkpoint-for-All" strategy and fail to generalize across the continuous reliability spectrum.

A.5 Computational Efficiency

To assess real-world feasibility, we evaluate the computational overhead on a single NVIDIA RTX 4090 GPU (Batch=16). Despite maintaining a model capacity of 112.47M parameters shown in Table 10, QA-MoE achieves a low computational cost of 4.33 GFLOPs per sample, attributed to the sparse activation of experts (only Top-3 are active). With an inference latency of 10.59 ms and a throughput of 1,510 samples/sec, our framework fully satisfies the requirements for real-time deployment.

Model	Parameters (M)		Comp.	Speed	Mem.
	Total	Active	GFLOPs ↓	Latency (ms) ↓	VRAM (MB) ↓
MulT	113.1	113.1	6.82	14.5	1350
MISA	114.4	114.4	5.95	12.8	1600
MMA	113.5	113.5	6.10	13.2	1420
SAM-LML	112.8	112.8	5.15	11.9	1100
QA-MoE (Ours)	113.2	**38.4***	**4.33**	**10.59**	**950**

Table 10: **Computational Efficiency Analysis.** We compare QA-MoE against strong baselines on a single RTX 4090 GPU (Batch=16). "Active" denotes the number of parameters actually used during a single inference pass.