Uncertainty-Calibrated Elastic Alignment for Multimodal Sentiment Analysis with Missing Modalities

Kang He[1,2], **Yuzhe Ding**[1], **Rao Fu**[1], **Yukang Feng**[2,3], **Kaipeng Zhang**[2,3], **Yiming Liu**[2], **Fei Li**[1], **Chong Teng**[1], **Donghong Ji**[1]*

[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University

[2]Shanghai Innovation Institute [3] Alaya Studio, Shanda AI Research Tokyo

{hekang0225, yuzheding, dhji}@whu.edu.cn

Abstract

Multimodal sentiment analysis (MSA) in real-world scenarios is often challenged by dynamically missing modalities. Existing methods predominantly rely on deterministic imputation and rigid alignment, which compels the model to overfit noise in ambiguous regions while neglecting the decision shift induced by modality inertia. To address these issues, we propose a novel uncertainty-calibrated elastic alignment framework, termed EASE. Specifically, we employ probabilistic imputation to capture cross-modal ambiguity and leverage the estimated uncertainty to drive elastic alignment, thereby adaptively relaxing constraints in ambiguous regions to avoid rigid fitting. Meanwhile, we introduce cross-view predictive consistency constraints to unify discriminative logic across different modality views, stabilizing the decision boundary under modality degradation. Extensive experiments demonstrate that EASE consistently outperforms existing state-of-the-art baselines across multiple benchmarks, exhibiting exceptional robustness particularly under high missing-rate scenarios.

1 Introduction

Multimodal Sentiment Analysis (MSA) leverages complementary semantics across linguistic, acoustic, and visual modalities (Yang et al., 2023; Wang et al., 2024; Zhou et al., 2025) to recognize human emotions (Gandhi et al., 2023; He et al., 2026a). However, in real-world scenarios, multimodal inputs are often dynamically incomplete (Li et al., 2024b; He et al., 2026b). Factors such as sensor malfunctions, environmental background noise, or privacy masking inevitably induce uncertainty-driven modality missingness. This often leads to catastrophic performance degradation in models trained on complete data (Zhang et al., 2024).

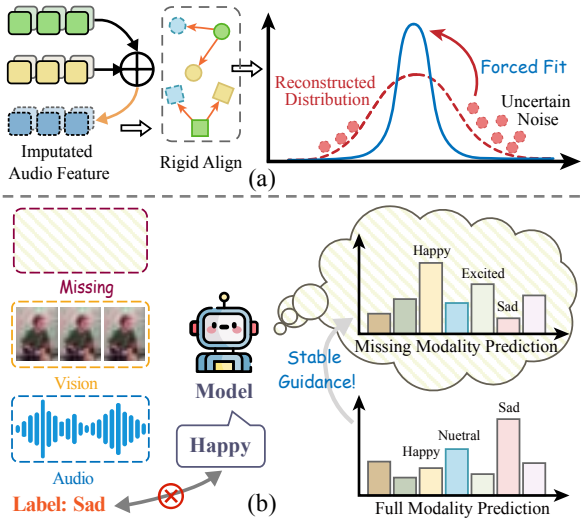


Figure 1: Illustration of modality representation and prediction challenges under missing modalities. (a) Forcing uncertain imputed features to match precise targets leads to a noisy **forced fit**. (b) Inconsistency between partial and full views causes **drastic fluctuations** in sentiment prediction (*e.g.*, Happy → Sad).

Recently, meaningful progress has been made in handling missing modalities via reconstruction (Yuan et al., 2021; Guo et al., 2024), feature alignment (Li et al., 2024c), and uncertainty modeling (Ma et al., 2021; Zhu et al., 2025), which effectively enhance robustness by restoring missing views or learning latent distributions. Nevertheless, prior works (Yuan et al., 2021) treat cross-modal completion as a deterministic one-to-one mapping, thereby overlooking the intrinsic uncertainty of missing modalities. Even when uncertainty (Lian et al., 2023; Zhu et al., 2025; Li et al., 2024a) is considered, conventional approaches often enforce rigid alignment constraints (*e.g.*, KL divergence), which compels the model to overfit stochastic errors as authentic semantics within high-uncertainty regions. Furthermore, these methods focus on feature-level fidelity while neglecting decision-level predictive consistency, failing to ensure stable discrimination under dynamic modality shifts.

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 5268–5288

Challenge I: The Rigid Alignment Dilemma. A key issue lies in how prior methods handle ambiguous cross-modal completion. The semantics of missing modalities possess *intrinsic ambiguity*. For instance, without acoustic intonation, the same textual utterance acts as a one-to-many mapping, potentially corresponding to multiple emotional tendencies. However, existing methods (Han et al., 2021; Zhang et al., 2024) typically ignore this stochasticity. As illustrated in Fig. 1(a), by employing point estimation coupled with rigid alignment, they forcibly fit a single target within these semantically ambiguous regions. Inevitably, the model treats stochastic reconstruction noise as authentic semantics. This not only distorts the feature space but also induces the learning of spurious correlations, thereby impairing classification performance.

Challenge II: Modality-Dependent Decision Shift. Moreover, a critical disconnect persists at the downstream decision level. As shown in Fig. 1(b), even if feature alignment appears to converge, dynamic view switching still triggers decision instability. Due to *modality inertia* (*i.e.*, an over-reliance on strong modalities) (Peng et al., 2022; Fan et al., 2023), the decision boundary drifts significantly when the input degrades from full-modality to partial-modality views. Existing methods (Wang et al., 2022a; Zhu et al., 2025) overlook the predictive consistency, failing to maintain invariant discriminative logic across different views. Consequently, the model struggles to establish a unified decision space, preventing completed features from translating into robust predictive capability.

To address these challenges, we propose a novel uncertainty-calibrated elastic alignment framework. We reformulate feature imputation as a probabilistic conditional density estimation problem, explicitly capturing the uncertainty inherent in ambiguous cross-modal mappings. Building on this, we introduce an uncertainty-adaptive elastic kernel in the Reproducing Kernel Hilbert Space (RKHS), which leverages the estimated confidence to modulate alignment strength and mitigate noise overfitting. Finally, we bridge the disconnect at the decision level by imposing consistency constraints between partial and full views, ensuring robust discrimination under varying missingness patterns. Our main contributions are summarized as follows:

- We propose EASE, shifting from fixed imputation to uncertainty-calibrated elastic adaptation for robust dynamic missingness.
- We introduce Uncertainty-Calibrated Semantic Alignment, which employs an uncertainty-adaptive elastic kernel to dynamically modulate alignment strength, effectively preventing rigid fitting in ambiguous regions.
- We design Modality-Invariant Predictive Consistency to stabilize decision boundaries by enforcing consistency across partial and full modalities, counteracting modality inertia.
- Extensive experiments on MOSI, MOSEI and SIMS demonstrate that EASE consistently outperforms SOTA methods, exhibiting superior robustness particularly in high-missingness scenarios.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) integrates text (He et al., 2025a,b), visual, and acoustic cues to recognize emotions. Prior works (Han et al., 2021; Yang et al., 2023; Fang et al., 2025a,b) enhance representation learning via noise-robust cross-modal interaction modeling and often uses language as a semantic anchor (Wu et al., 2021; Zhang et al., 2023a). For real-world dynamic missingness, existing methods mainly rely on reconstruction (Yuan et al., 2021) or distillation (Li et al., 2024c), with additional reliability modeling (Zhang et al., 2024) and prompt-based adaptation (Guo et al., 2024) to handle noise (Zhu et al., 2025; Li et al., 2025a). Despite these advances, most methods depend on deterministic reconstruction and rigid alignment, which over-constrain ambiguous regions and fail to ensure cross-view decision stability. Meanwhile, large language models (Sun et al., 2025b,a; Ai et al., 2025; Feng et al., 2025, 2026; Li et al., 2025b)s have been extended to multimodal sentiment analysis (Cheng et al., 2024), yet they still implicitly assume complete inputs and lack explicit uncertainty quantification under missing modalities. We propose EASE, which combines uncertainty-driven elastic alignment with predictive consistency constraints for robust inference.

2.2 Multimodal Alignment Learning

Multimodal alignment learning constructs a semantic bridge that maps heterogeneous modalities into a shared representation space. Beyond instance-level matching, recent works increasingly adopt

geometric and distribution-level objectives, including mutual information maximization (Dufumier et al., 2025), optimal transport (Chen et al., 2024; Rho et al., 2025), and hypergraph learning for high-order relations (Gu and Wang, 2025). For incomplete inputs, existing methods typically rely on feature generation (Dai et al., 2025), retrieval augmentation (Pipoli et al., 2025), and parameter-efficient adaptation (Reza et al., 2025). However, they often enforce alignment as a deterministic hard constraint, pulling unreliable imputations too aggressively. In contrast, we propose elastic alignment that calibrates alignment strength with uncertainty, thereby avoiding rigid constraints under inherent cross-modal ambiguity.

3 Methodology

Fig. 2 illustrates the key components and workflow of our proposed EASE framework. EASE performs probabilistic cross-modal completion, calibrates alignment strength via uncertainty, and enforces cross-view predictive consistency to stabilize decisions across partial and full modality views.

3.1 Problem Formulation

Consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i = \{x_i^m\}_{m\in\{t,v,a\}}$ denotes the aligned multimodal features and y_i the sentiment label. To simulate dynamic missingness, we apply a random mask M_i to generate incomplete views $\tilde{\mathbf{x}}_i = \mathrm{Mask}(\mathbf{x}_i; M_i)$. Our goal is to learn a mapping $f : \tilde{\mathbf{x}}_i \to y_i$ robust to varying missingness patterns.

3.2 Uncertainty-Aware Cross-Modal Imputation (UCMI)

Traditional imputation methods typically rely on deterministic point estimation, thereby overlooking the intrinsic ambiguity of cross-modal mapping (*e.g.*, a single textual utterance may correspond to diverse acoustic tones). Inspired by uncertainty estimation (Wang et al., 2022b; Hao and Zhang, 2023), we propose Uncertainty-Aware Cross-Modal Imputation (UCMI), which reformulates the imputation task as a conditional density estimation problem.

Given an available source modality x_A and a missing target modality x_B, we first extract the modality-specific feature z_A via a Transformer backbone (Vaswani et al., 2017). To bridge the modality gap, we project z_A into a common semantic space via a learnable projection head, yielding the context representation h_{ctx}. Instead of generating a deterministic vector $\hat{z}_B$, we employ h_{ctx} to model the latent target as a conditional Gaussian distribution $q_\phi(z_B|h_{\text{ctx}})$ and predict its parameters:

$$\boldsymbol{\mu}_B = \mathcal{F}_\mu(h_{\text{ctx}}) \tag{1}$$

$$\boldsymbol{\sigma}_B^2 = \text{Softplus}(\mathcal{F}_\sigma(h_{\text{ctx}})) + \epsilon \tag{2}$$

where $\mathcal{F}_\mu$ and $\mathcal{F}_\sigma$ are projection heads, yielding the completion distribution:

$$q_\phi(z_B|\tilde{\mathbf{x}}) = \mathcal{N}(\boldsymbol{\mu}_B, \mathrm{diag}(\boldsymbol{\sigma}_B^2)) \tag{3}$$

Here, the mean $\boldsymbol{\mu}_B$ captures representative semantics, while the variance $\boldsymbol{\sigma}_B^2$ reflects *the unique information* of the missing modality that is not explained by the observed ones.

During training, we supervise this probabilistic completion process using the ground truth features z_B^{gt} corresponding to the complete samples. We optimize by minimizing the negative log-likelihood of the ground truth features under the predicted distribution, detailed in Appendix B.1:

$$\begin{aligned}\mathcal{L}_{\text{ucmi}} &= -\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\log q_\phi(z_B^{gt} \mid \tilde{\mathbf{x}})\right] \\ &= \frac{1}{2}\sum_{d=1}^{D}(\log(2\pi\sigma_{B,d}^2) + \frac{(z_{B,d}^{gt} - \mu_{B,d})^2}{\sigma_{B,d}^2})\end{aligned} \tag{4}$$

Notably, while P-RMF (Zhu et al., 2025) uses Gaussian embeddings mainly for feature fusion, we leverage the learned variance $\boldsymbol{\sigma}_B^2$ as a **calibration signal**. It is further propagated to modulate alignment strength, preventing noise from high-uncertainty regions from contaminating downstream representations.

3.3 Uncertainty-Calibrated Semantic Alignment (UCSA)

Generated features under missing-modality settings exhibit heterogeneous uncertainty. However, standard objectives such as MSE (Lian et al., 2023; Zhu et al., 2025) or KL divergence (Li et al., 2024a) impose uniform, rigid constraints, forcing the model to fit reconstruction targets regardless of reliability.

To remedy this, we propose Uncertainty-Calibrated Semantic Alignment (UCSA), which leverages uncertainty as a reliability cue to adaptively modulate alignment strength. UCSA enables elastic alignment at both the *distributional level* and the *class-semantic level*, preventing over-constraining in ambiguous regions while preserving semantic consistency.

Uncertainty-Adaptive Distributional Alignment. Given the completed feature $\hat{z}$ and sample-level uncertainty $u_i = \frac{1}{D}\|\boldsymbol{\sigma}_i^2\|_1$, we align the completed

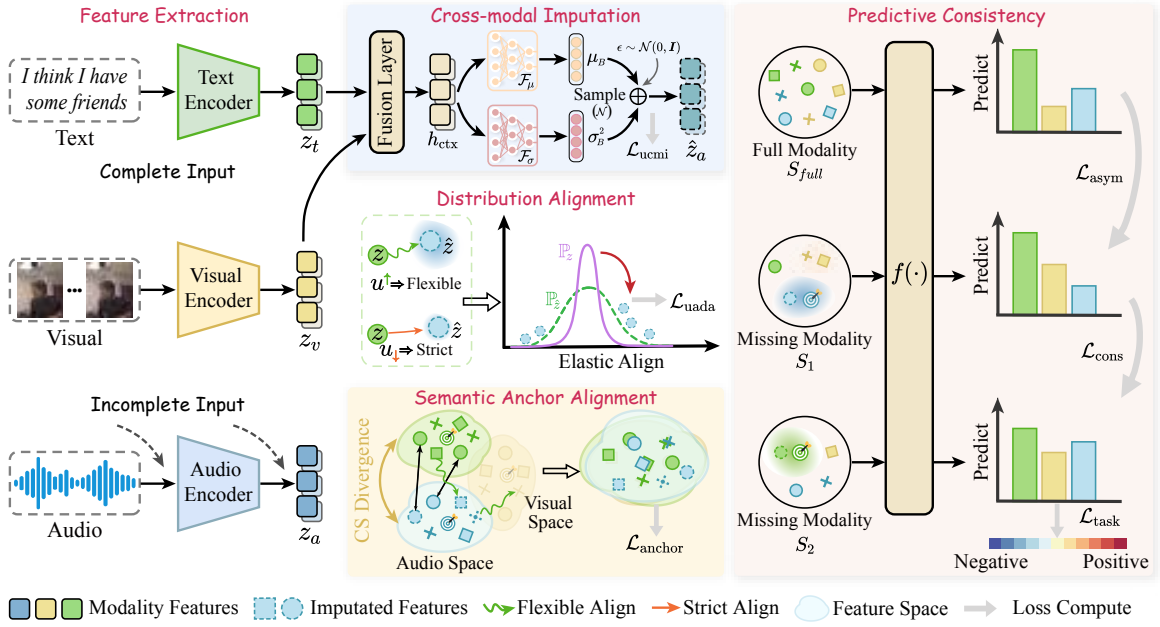


Figure 2: The overall architecture of our proposed model EASE.

feature distributions $\mathbb{P}_{\hat{z}}$ and the real feature distribution $\mathbb{P}_z$ via a novel *uncertainty-adaptive kernel* in the Reproducing Kernel Hilbert Space (RKHS) (Berlinet and Thomas-Agnan, 2011):

$$\mathcal{K}_{ua}(z_i, \hat{z}_j) = \exp\left(-\frac{\|z_i - \hat{z}_j\|^2}{2\tau^2 \cdot \psi(u_i, \hat{u}_j)}\right) \tag{5}$$

We regulate the alignment receptive field via a dynamic bandwidth ψ:

$$\psi(u_i, \hat{u}_j) = 1 + \alpha \cdot \text{Mean}(u_i, \hat{u}_j) \tag{6}$$

Specifically, higher uncertainty ($u \uparrow$) triggers bandwidth expansion, widening the admissible neighborhood for generated features, whereas lower uncertainty induces contraction to enforce stricter alignment. This shifts the optimization objective from rigid point-to-point matching to elastic distributional overlapping. Based on this adaptive kernel, we minimize the Cauchy-Schwarz (CS) divergence (Principe et al., 2000; Yin et al., 2025):

$$\mathcal{L}_{\text{uada}} = -\log\left(\frac{\sum_{i,j}\mathcal{K}_{ua}(z_i, \hat{z}_j)}{\sqrt{\sum_{i,j}\mathcal{K}_{ua}(z_i, z_j)\cdot\sum_{i,j}\mathcal{K}_{ua}(\hat{z}_i, \hat{z}_j)}}\right) \tag{7}$$

Semantic Anchor Alignment. Given that boundaries between different sentiment classes are inherently fuzzy, relying solely on distributional alignment is insufficient to ensure that generated features maintain consistency within class semantics (Wen et al., 2016). Therefore, we explicitly model the class-conditional feature distribution to provide a stable semantic reference.

We explicitly model each sentiment class as a Gaussian anchor $\mathcal{P}_k = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and constrain the completion distribution $\mathcal{Q} = \mathcal{N}(\hat{\boldsymbol{z}}, \hat{\boldsymbol{\Sigma}})$ to converge towards its target. The detailed derivation is given in Appendix B.2. The objective is derived from the analytical CS divergence:

$$\mathcal{L}_{\text{anchor}} = D_{CS}(\mathcal{Q}\|\mathcal{P}_y) = \underbrace{\log\frac{|\hat{\boldsymbol{\Sigma}} + \boldsymbol{\Sigma}_y|}{\sqrt{|2\hat{\boldsymbol{\Sigma}}|\cdot|2\boldsymbol{\Sigma}_y|}}}_{\text{Distribution Shape Term}} + \underbrace{(\hat{\boldsymbol{z}} - \boldsymbol{\mu}_y)^\top(\hat{\boldsymbol{\Sigma}} + \boldsymbol{\Sigma}_y)^{-1}(\hat{\boldsymbol{z}} - \boldsymbol{\mu}_y)}_{\text{Semantic Centrality Term}} \tag{8}$$

Notably, the inverse covariance term $(\hat{\boldsymbol{\Sigma}}+\boldsymbol{\Sigma}_y)^{-1}$ functions as an adaptive gating mechanism: high uncertainty inherently attenuates the penalty on the semantic distance. In synergy with uncertainty-adaptive distributional alignment, this ensures elastic semantic consistency. The overall alignment objective is formulated as:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{uada}} + \beta_1\mathcal{L}_{\text{anchor}} \tag{9}$$

where β_1 is hyperparameter.

3.4 Modality-Invariant Predictive Consistency

Recent studies (Wei et al., 2024; Fan et al., 2023, 2024) reveal that multimodal models often exhibit an over-reliance on dominant modalities. When inputs degrade from full to partial views, this dependency is disrupted, causing the classifier to deviate from established decision paths and leading to significant predictive instability (Ma et al., 2021;

Zhang et al., 2024). To address this, we propose Modality-Invariant Predictive Consistency (MIPC), which enforces consistent predictive distributions across views, thereby decoupling the model's dependence on specific modality combinations.

Symmetric Modality Consistency. Let $\mathbf{p}_i^{(S)} = f(\mathbf{z}_i^{(S)})$ denote the output probability distribution of the classifier for a modality subset S. To ensure the stability of sentiment prediction under different missing patterns, we constrain any two views $S_1, S_2 \subseteq \{t, a, v\}$ of the same sample to maintain predictive consistency:

$$\mathcal{L}_{\text{cons}} = \mathbb{E}_{S_1,S_2}\left[\text{KL}\big(\mathbf{p}^{(S_1)} \parallel \mathbf{p}^{(S_2)}\big)\right] \tag{10}$$

This constraint penalizes drastic fluctuations in the prediction distribution during view switching, thereby counteracting the boundary drift caused by dynamic missingness at the decision level.

Asymmetric Distillation via Full-Modality Anchoring. Considering that the full-modality view S_{full} typically contains the most complete semantic information, we treat it as a highly reliable semantic anchor. Through an asymmetric constraint, we guide any incomplete view S_p towards the full-modality decision boundary:

$$\mathcal{L}_{\text{asym}} = \mathbb{E}_{S \subset S_{\text{full}}}\left[\text{KL}\big(\mathbf{p}^{(S)} \parallel \text{sg}(\mathbf{p}^{(S_{\text{full}})})\big)\right] \tag{11}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. This mechanism effectively injects stable decision knowledge from the full view into the partial views laden with uncertainty. The final consistency regularization term is:

$$\mathcal{L}_{\text{consist}} = \mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{asym}} \tag{12}$$

3.5 Optimization Objective

The EASE framework is trained jointly in an end-to-end manner. The total optimization objective is composed of the task prediction loss and the aforementioned losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{ucmi}} + \lambda_2 \mathcal{L}_{\text{align}} + \lambda_3 \mathcal{L}_{\text{consist}} \tag{13}$$

where $\mathcal{L}_{\text{task}}$ is the standard cross-entropy loss, λ_1 controls the strength of the imputation supervision, λ_2 weights the semantic alignment regularization, and λ_3 scales the predictive consistency constraint.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments and analysis on three widely used benchmarks: MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and SIMS (Yu et al., 2020). Detailed dataset descriptions are provided in Appendix A.1.

Evaluation Metrics. We employ Acc-2, F1 scores, MAE, and Corr across all datasets, supplemented by Acc-7 for MOSI/MOSEI and Acc-3/Acc-5 for SIMS. For binary metrics (Acc-2, F1 scores), we report results for both non-negative and positive class definitions (denoted as '-/-'). Detailed definitions are provided in Appendix A.2.

Implementation Details. Final performance is averaged over all missing rates. Models are trained for 200 epochs with a batch size of 64 on a single NVIDIA RTX 4090 GPU. More details are provided in Appendix A.3.

4.2 Baselines

We conduct a fair and comprehensive comparison against state-of-the-art baselines, which cover complete-modality methods, including MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), CENET (Wang et al., 2022a), TETFN (Wang et al., 2023), and ALMT (Zhang et al., 2023b), as well as missing-modality methods, including TFR-Net (Yuan et al., 2021), LNLN (Zhang et al., 2024), P-RMF (Zhu et al., 2025), and TF-Mamba (Li et al., 2025a). Baseline details are provided in Appendix A.5.

4.3 Robust Comparison

Robustness Analysis against Intra-modal Missingness. To evaluate robustness against intra-modality missingness, we apply random masking with missing rates $\{0, 0.1, \ldots, 0.9\}$ and report the average performance across all rates. Tables 1 and 2 summarize the results on MOSI, MOSEI, and SIMS. Overall, EASE achieves state-of-the-art (SOTA) or highly competitive performance across multiple metrics.

As shown in Table 1, on the MOSI, EASE achieves an Acc-2 of 73.60%, yielding a 1.46% relative improvement over the strong baseline TF-Mamba (72.54%). Moreover, EASE consistently improves regression metrics, reducing MAE from 1.035 to 0.995 (3.86%) and increasing Corr from 0.548 to 0.573 (4.56%). These results demonstrate that leveraging variance signals to suppress unreliable imputed features effectively mitigates noise from missing data and enhances predictive stability. Full results are provided in Appendix C.1.

Model	MOSI						MOSEI					
	Acc-2↑	F1↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑	Acc-2↑	F1↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
MISA	70.33/71.49	70.00/71.28	33.08	29.85	1.085	0.524	75.82/71.27	68.73/63.85	39.39	40.84	0.780	0.503
Self-MM	69.26/70.51	67.54/66.60	34.67	29.55	1.070	0.512	77.42/73.89	72.31/68.92	45.38	44.70	0.695	0.498
MMIM	67.06/69.14	64.04/66.65	33.77	31.30	1.077	0.507	75.89/73.32	70.32/68.72	41.74	40.75	0.739	0.489
CENET	67.73/71.46	64.85/68.41	37.25	30.38	1.080	0.504	77.34/74.67	74.08/70.68	47.83	47.18	0.685	0.535
TETFN	67.68/69.76	63.29/65.69	34.34	30.30	1.087	0.507	67.68/69.76	63.29/65.69	47.70	30.30	1.087	0.508
TFR-Net	66.35/68.15	60.06/61.73	34.67	29.54	1.200	0.459	77.23/73.62	71.99/68.80	34.67	46.83	0.697	0.489
ALMT	68.39/70.40	71.80/72.57	33.42	30.30	1.083	0.498	77.54/76.64	78.03/77.14	41.64	40.92	0.674	0.481
LNLN	70.94/72.55	71.25/72.73	38.27	34.26	1.046	0.527	78.19/76.30	79.95/77.77	46.17	45.42	0.692	0.530
P-RMF	71.53/72.81	71.69/72.93	38.50	34.19	1.038	0.525	78.83/78.14	**80.39**/79.33	45.87	44.63	0.658	0.589
TF-Mamba	72.54/73.46	72.57/73.59	37.74	33.95	1.035	0.548	77.61/77.34	77.43/77.18	46.64	45.66	0.673	0.578
EASE	**73.60/74.71**	**73.66/74.79**	**39.57**	**35.75**	**0.995**	**0.573**	**79.43/79.14**	80.04/**79.67**	**47.69**	**46.17**	**0.645**	**0.610**

Table 1: Robustness comparison of the overall performance on the MOSI and MOSEI datasets under intra-modal missingness, simulated via random masking with missing rates $\{0, 0.1, \ldots, 0.9\}$. Refer to Sec. 4.3.

Model	Acc-2↑	F1↑	Acc-3↑	Acc-5↑	MAE↓	Corr↑
MISA	72.71	66.30	56.87	31.53	0.539	0.348
Self-MM	72.81	68.43	56.75	32.28	0.508	0.376
MMIM	69.86	66.21	52.76	31.81	0.544	0.339
TFR-Net	68.13	58.70	52.89	26.52	0.661	0.169
CENET	68.13	57.90	53.17	22.29	0.589	0.107
ALMT	69.66	72.76	45.36	20.00	0.561	0.364
LNLN	72.73	**79.43**	57.14	34.64	0.514	0.397
P-RMF	73.64	74.65	54.75	34.83	0.500	0.414
TF-Mamba	74.68	72.20	55.51	34.46	0.512	0.386
EASE	**74.72**	75.13	**58.08**	**35.62**	**0.486**	**0.451**

Table 2: Robustness comparison on the SIMS dataset under intra-modal missingness. Details in Sec. 4.3.

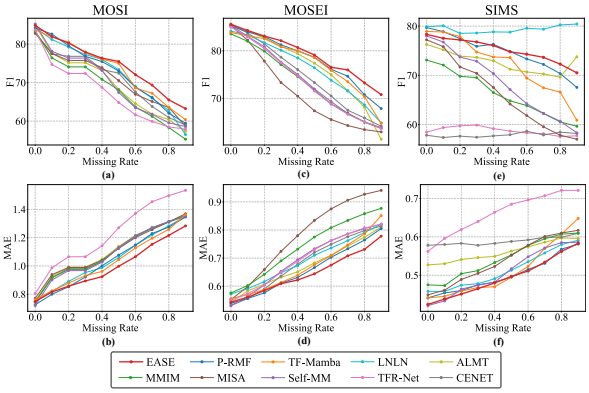


Figure 3: Performance curves under varying random missing rates across MOSI, MOSEI, and SIMS datasets. Best viewed in color. Please refer to Sec. 4.3.

As detailed in Table 2, EASE demonstrates significant improvements on the SIMS dataset. Compared to TF-Mamba (Li et al., 2025a), EASE improves F1 and Acc-3 by 4.05% and 4.63%, respectively, validating its robustness under varying noise intensities and missingness conditions. Notably, EASE does not achieve the highest F1 score on SIMS. However, observing Fig. 3(e) and (f) reveal an anomaly: the F1 score of LNLN(Zhang et al., 2024) counter-intuitively rises as the missing rate increases, while its MAE fails to show a corresponding reduction. This suggests that LNLN exhibits bias in missing scenarios; the model likely maintains superficial performance by defaulting to the majority class, rather than learning stable, cross-view predictive patterns.

Fig. 3 illustrates the performance trajectories across varying missing rates. While performance universally degrades as missingness increases, EASE (*red line*) exhibits the most graceful degradation trend. Even in high-missingness intervals ($MR \geq 0.6$), EASE maintains a superior performance floor (*e.g.*, F1 on MOSEI and SIMS), underscoring its capability to robustly model the uncertainty introduced by missing modalities.

Robustness Analysis against Inter-modal Missingness. Following prior studies (Zhang et al., 2024; Li et al., 2025a), we conduct experiments under conditions where only partial modality subsets $m \in \{l\}/\{a\}/\{v\}/\{l, a\}/\{l, v\}/\{a, v\}$ are retained as inputs, and report the averaged results across these settings.

As shown in Table 3, EASE achieves state-of-the-art or competitive performance across the majority of metrics, demonstrating robust generalization capabilities. On MOSI, compared to the latest baseline P-RMF, EASE improves Acc-2 and Acc-5 by 3.86% and 6.16%, respectively. Furthermore, it attains the best MAE and Corr, indicating that EASE maintains reliable regression consistency even when key modalities are absent. Similarly, on MOSEI, EASE achieves superior performance in metrics such as Acc-5, MAE, and Corr.

We note that although P-RMF (Zhu et al., 2025) and LNLN exhibit a slight F1 advantage, their elevated MAE and Corr suggest a *suboptimal trade-off*: overfitting classification boundaries at the expense of the fidelity of fine-grained sentiment inten-

Model	MOSI						MOSEI					
	Acc-2↑	F1↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑	Acc-2↑	F1↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
MISA	67.28/67.64	64.31/64.81	31.77	29.44	1.096	0.462	75.49/73.94	71.31/66.65	40.32	41.45	0.752	0.438
Self-MM	68.17/70.75	61.91/63.71	35.61	29.75	1.052	0.450	75.86/74.12	71.82/67.00	46.06	42.67	0.694	0.447
MMIM	65.12/68.35	58.28/63.38	33.74	31.28	1.059	0.444	75.45/73.19	70.18/68.12	42.54	41.36	0.733	0.449
CENET	66.67/70.38	58.61/62.66	33.95	30.30	1.069	0.451	74.15/73.45	70.37/68.44	43.19	42.41	0.694	0.470
TETFN	68.15/70.17	60.18/62.47	35.50	32.67	1.061	0.446	75.74/73.87	71.46/66.71	43.42	42.53	0.690	0.429
TFR-Net	67.96/70.50	60.09/62.88	34.97	29.77	1.177	0.451	75.92/73.51	71.28/66.89	43.09	42.33	0.701	0.427
ALMT	67.47/70.53	73.35/75.90	30.74	27.87	1.130	0.447	63.38/60.66	71.24/70.64	27.46	25.94	0.711	0.355
LNLN	65.64/68.53	69.66/72.02	34.86	31.94	1.094	0.424	**77.05**/73.53	**83.20**/80.85	44.50	43.86	0.733	0.425
P-RMF	68.14/69.41	**76.21/76.72**	34.45	31.64	1.072	0.433	76.81/73.34	82.76/79.62	44.30	42.67	0.696	0.469
TF-Mamba	69.31/68.51	68.57/67.65	33.58	29.78	1.147	0.439	72.59/75.00	69.90/71.65	44.91	44.20	0.738	0.439
EASE	**70.45/72.10**	73.99/76.11	**37.17**	**33.59**	**1.035**	**0.473**	76.86/**75.79**	82.81/**80.96**	**46.95**	**44.77**	**0.682**	**0.487**

Table 3: Performance comparison on the MOSI and MOSEI datasets under inter-modal missingness, including modality settings $m \in \{l\}/\{a\}/\{v\}/\{l,a\}/\{l,v\}/\{a,v\}$. Please see details in Sec. 4.3.

Model	MOSI		MOSEI		SIMS	
	F1↑	Acc-5↑	F1↑	Acc-5↑	F1↑	Acc-5↑
Ablation Study						
EASE	**73.66/74.79**	**39.57**	**80.04/79.67**	**47.69**	**75.13**	**35.62**
w/o UCMI	72.15/72.98	37.26	78.41/78.07	45.34	73.25	33.29
w/o UCSA	72.80/73.84	38.65	79.12/78.90	46.58	74.29	34.87
w/o MIPC	72.67/73.41	38.02	78.89/78.74	46.13	74.06	34.54
w/o ALL	71.47/72.38	36.53	77.65/78.09	44.67	72.70	32.60
Ablation on Loss Components						
EASE	**73.66/74.79**	**39.57**	**80.04/79.67**	**47.69**	**75.13**	**35.62**
w/o $\mathcal{L}_{\text{uada}}$	73.14/74.23	39.06	79.51/79.34	46.90	74.67	35.14
w/o $\mathcal{L}_{\text{anchor}}$	72.95/73.96	38.81	79.29/79.03	46.68	74.40	35.01
w/o $\mathcal{L}_{\text{cons}}$	73.15/73.92	38.77	79.43/78.88	46.73	74.45	35.06
w/o $\mathcal{L}_{\text{asym}}$	72.88/73.68	38.40	79.11/79.06	46.39	74.12	34.84

Table 4: Ablation studies of the key componets and different loss terms in EASE. Refer to Sec. 4.4.

sity prediction. In contrast, EASE strikes a more **effective balance between classification performance and regression stability**. Full results are provided in Appendix C.2.

4.4 Ablation Studies

To verify the effectiveness of each component, we conduct a comprehensive ablation study. As summarized in Table 4, removing any individual module within EASE results in a degradation of overall performance. Specifically, w/o UCMI causes the most severe deterioration across all datasets (*e.g.*, a 1.81% drop in Acc-2 on MOSI), validating the necessity of utilizing variance signals to suppress noise interference. Similarly, the absence of UCSA and MIPC leads to consistent performance declines, providing strong empirical support for the effectiveness of elastic alignment and predictive consistency in enhancing noise adaptation capabilities. The substantial gap between EASE and the baseline backbone ('w/o ALL') further attests to the synergistic efficacy of the overall architecture.

We further dissect the contributions of **specific loss terms**. Both distributional alignment ($\mathcal{L}_{\text{uada}}$) and anchor constraints ($\mathcal{L}_{\text{anchor}}$) prove to be indispensable. Notably, removing results in a more pronounced drop in F1 scores, suggesting that class-conditional anchors effectively prevent feature drift in high-uncertainty regions induced by missing modalities. Furthermore, the exclusion of the asymmetric constraint ($\mathcal{L}_{\text{asym}}$) leads to a significant decline in Acc-5, highlighting the necessity of leveraging the full-modality view as a reliable supervisor to guide the learning of incomplete views.

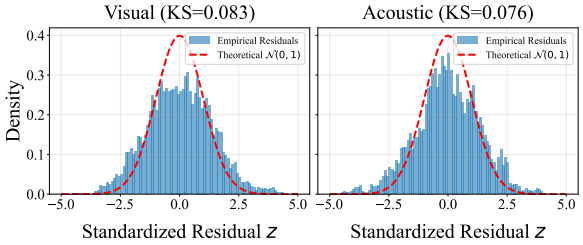


Figure 4: Statistical validation of uncertainty estimation on the MOSEI test set. Refer to Sec. 4.5.

As shown in Table 6, **single-modality ablations** confirm that while all modalities contribute positively. The removal of language causes the most severe degradation, highlighting the inherent ambiguity of audio-visual signals in constraining semantic inference, which is consistent with Fig. 5. Please see details in Appendix C.3.

4.5 In-depth Analysis

Uncertainty Estimation Verification. To ensure the predicted variance is a reliable indicator of missing modality uncertainty, we examine the distribution of standardized residuals. Ideally, a well-calibrated model produces residuals that follow $\mathcal{N}(0,1)$. As shown in Fig. 4, the residual distributions for both acoustic and visual modalities largely overlap with the standard normal density, suggesting that the model is neither over-confident nor under-confident. Quantitatively, the robustness of

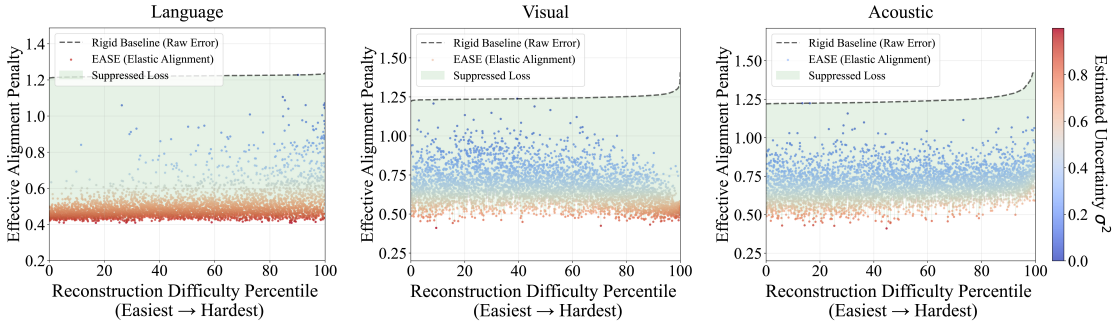
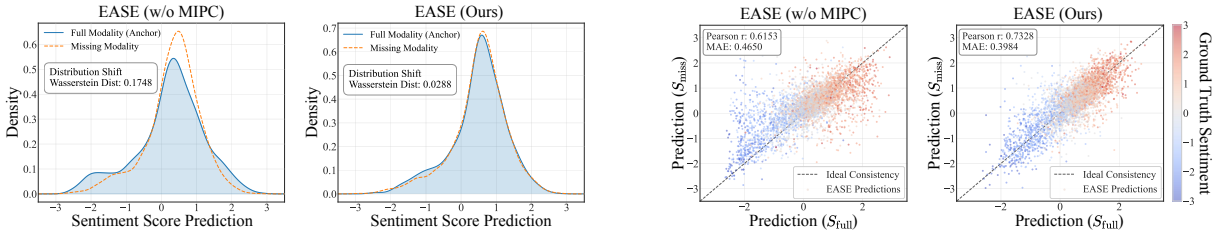


Figure 5: Elastic Alignment Visualization on MOSEI. Sorted by reconstruction difficulty, the plot contrasts the rigid baseline (dashed) with EASE (points), where vertical drops indicate uncertainty-driven penalty. Refer to Sec. 4.5.



(a) Distribution Alignment

(b) Sentiment Predictive Consistency

Figure 6: Visualization of the MIPC module's impact on (a) distribution-level stability and (b) instance-level predictive consistency on the MOSEI dataset. Best viewed in color. Refer to Sec. 4.5 for details.

this alignment is further confirmed by the low Kolmogorov–Smirnov statistics of 0.076 and 0.083, respectively. These metrics support the validity of the Gaussian assumption underlying our UCMI module and confirm that *the generated variance is statistically meaningful*. We provide a further discussion on the theoretical suitability in Appendix B.1.

Effect of Elastic Alignment. To verify the efficacy of elastic alignment, we visualize the effective alignment penalty versus reconstruction difficulty in Fig. 5 (Appendix A.6). Three insights emerge: ❶ The rigid baseline (dashed) *retains a high penalty floor even for easy samples*, indicating substantial irreducible noise in multimodal data. ❷ EASE **globally relaxes alignment strength** via the dynamic bandwidth ψ (green shaded region), acting as a soft filter that reduces overfitting to ambiguous or noisy residuals. ❸ For the Language modality, EASE sharply lowers the penalty to ~ 0.4, reflecting adaptive recognition of *high aleatoric uncertainty* when inferring text from audio-visual cues (Appendix A.6). This relaxation prevents fitting non-informative noise while preserving semantic consistency. We conduct the same experiments on MOSI, with detailed results in Appendix C.4.

Visualization of Predictive Consistency. To further validate MIPC, we conduct a multi-level visualization analysis on MOSEI (Fig. 6). As illustrated in Fig. 6a, MIPC *aligns the prediction distributions* of completed and observed inputs, significantly reducing the Wasserstein Distance (0.1748 $\rightarrow$ 0.0288). This confirms that MIPC effectively suppresses distribution shifts induced by missing modalities. At the instance level (Fig. 6b), MIPC promotes *tighter clustering around the diagonal*, increasing the Pearson correlation (0.6153 $\rightarrow$ 0.7328) while improving MAE to 0.398. Collectively, these results indicate that MIPC not only enhances stability in a statistical distributional sense but also rigorously constrains predictive discrepancies between completed and observed views at the instance level.

Model	Params	Time/Epoch	Acc-7
SelfMM	122M	42.6s	45.38
LNLN	116M	24.0s	46.17
P-RMF	117M	18.0s	45.87
EASE	116M	7.9s	47.69

Table 5: Comparison of parameters, computational efficiency and Acc-7 performance on the MOSI dataset across different baselines.

Efficiency and Sensitivity Analysis. We further evaluate the computational efficiency and hyperparameter sensitivity of EASE. As shown in Table 5, EASE achieves the best Acc-7 (47.69) on MOSI with 116.5M parameters, comparable to LNLN (116M) and P-RMF (117M), while requiring only 7.9s per epoch—a $\sim$3.0$\times$ speedup over LNLN

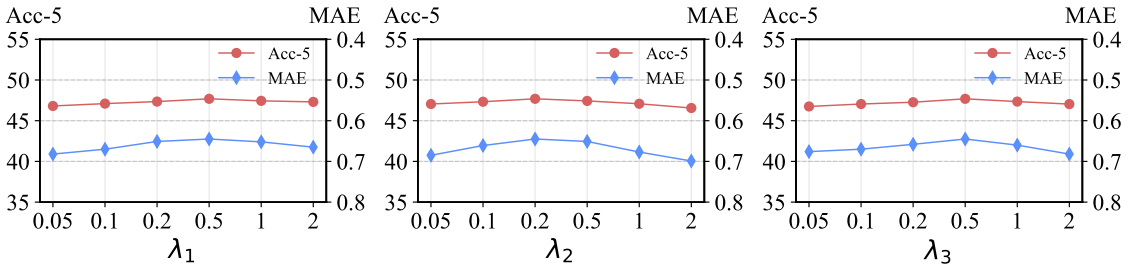


Figure 7: Sensitivity analysis on the MOSEI dataset.

and $\sim$5.4$\times$ over SelfMM, demonstrating a favorable efficiency–performance trade-off. For sensitivity, we independently vary each loss weight in Eq. 13 across $\{0.05, 0.1, 0.2, 0.5, 1, 2\}$. As shown in Fig. 7, both Acc-5 and MAE on MOSEI remain stable within $\lambda \in [0.1, 1]$, confirming that EASE does not rely on fragile hyperparameter tuning. The distinct optima (λ_1=0.5, λ_2=0.2, λ_3=0.5) suggest that imputation and consistency benefit from moderately strong supervision, while elastic alignment should be applied more gently to avoid over-constraining uncertain features.

5 Conclusion

In this paper, we proposed EASE to address the challenges of intrinsic feature ambiguity and decision instability in real-world MSA. By modeling missingness as a probabilistic problem, our framework enables uncertainty-guided elastic alignment, which adaptively relaxes constraints in ambiguous regions to avoid noise overfitting. Additionally, we enforce cross-view predictive consistency to stabilize decision boundaries against dynamic modality shifts. Extensive experiments on benchmark datasets demonstrate that EASE effectively mitigates performance degradation under uncertainty, consistently outperforming SOTA methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62176187).

Limitations

Despite the improvements of EASE, several limitations remain. First, due to computational constraints, we referenced baselines from recent benchmarks (Zhang et al., 2024) rather than re-implementing all methods, following identical splits and protocols for fairness (Appendix C.1). Second, EASE's gains are less pronounced under low missingness, where uncertainty calibration overhead may outweigh its benefits (Appendix C.1). Third, our method targets offline settings and does not handle streaming or temporally correlated missing patterns, which we leave for future work. Finally, while EASE is evaluated on multimodal sentiment analysis, its applicability to other multimodal understanding tasks, such as stance detection (Ding et al., 2025), remains unexplored. We leave the extension to these broader tasks and online scenarios for future work.

References

Jiaxin Ai, Pengfei Zhou, Zhaopan Xu, Ming Li, Fanrui Zhang, Zizhen Li, Jianwen Sun, Yukang Feng, Baojin Huang, Zhongyuan Wang, and Kaipeng Zhang. 2025. Projudge: A multi-modal multi-discipline benchmark and instruction-tuning dataset for mllm-based process judges. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4681–4690.

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.

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. IEEE.

Alain Berlinet and Christine Thomas-Agnan. 2011. *Reproducing kernel Hilbert spaces in probability and statistics*.

Zihao Chen, Chi-Heng Lin, Ran Liu, Jingyun Xiao, and Eva L. Dyer. 2024. Your contrastive learning problem is secretly a distribution alignment problem. In *Advances in Neural Information Processing Systems*, pages 91597–91617.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. 2025. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24507–24517.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Yuzhe Ding, Kang He, Bobo Li, Li Zheng, Haijun He, Fei Li, Chong Teng, and Donghong Ji. 2025. Zero-shot conversational stance detection: Dataset and approaches. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3221–3235.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. 2025. What to align in multimodal contrastive learning? In *The Thirteenth International Conference on Learning Representations*.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junhong Liu, and Song Guo. 2024. Detached and interactive multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5470–5478.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. 2025a. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14314–14324.

Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. 2025b. Catch your emotion: Sharpening emotion perception in multimodal large language models. In *Forty-second International Conference on Machine Learning*.

Yukang Feng, Jianwen Sun, Chuanhao Li, Zizhen Li, Jiaxin Ai, Fanrui Zhang, Yifan Chang, Sizhuo Zhou, Shenglin Zhang, Yu Dai, and 1 others. 2025. A high-quality dataset and reliable evaluation for interleaved image-text generation. *arXiv preprint arXiv:2506.09427*.

Yukang Feng, Jianwen Sun, Zelai Yang, Jiaxin Ai, Chuanhao Li, Zizhen Li, Fanrui Zhang, Kang He, Rui Ma, Jifan Lin, and 1 others. 2026. Longcli-bench: A preliminary benchmark and study for long-horizon agentic programming in command-line interfaces. *arXiv preprint arXiv:2602.14337*.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Zhibin Gu and Weili Wang. 2025. Hypergraph-enhanced contrastive learning for multi-view clustering with hyper-laplacian regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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.

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.

Xiaoshuai Hao and Wanqian Zhang. 2023. Uncertainty-aware alignment network for cross-domain video-text retrieval. *Advances in Neural Information Processing Systems*, 36:38284–38296.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th Association for Computing Machinery International Conference on Multimedia*, pages 1122–1131.

Kang He, Boyu Chen, Yuzhe Ding, Fei Li, Chong Teng, and Donghong Ji. 2026a. Pase: Prototype-aligned calibration and shapley-based equilibrium for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30960–30968.

Kang He, Yuzhe Ding, Bobo Li, Haining Wang, Fei Li, Chong Teng, and Donghong Ji. 2025a. Harnessing dimensional contrast and information compensation for sentence embedding enhancement. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Kang He, Yuzhe Ding, Haining Wang, Fei Li, Chong Teng, and Donghong Ji. 2025b. DALR: Dual-level alignment learning for multimodal sentence representation learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3586–3601.

Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. 2026b. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis. *arXiv preprint arXiv:2604.12518*.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. 2024a. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10074–10082.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, and 1 others. 2024b. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. *Advances in Neural Information Processing Systems*, 37:28515–28536.

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

Xiang Li, Xianfu Cheng, Dezhuang Miao, Xiaoming Zhang, and Zhoujun Li. 2025a. Tf-mamba: Text-enhanced fusion mamba with missing modalities for robust multimodal sentiment analysis. *arXiv preprint arXiv:2505.14329*.

Zizhen Li, Chuanhao Li, Yibin Wang, Qi Chen, Diping Song, Yukang Feng, Jianwen Sun, Jiaxin Ai, Fanrui Zhang, Mingzhu Sun, and 1 others. 2025b. Inmind: Evaluating llms in capturing and applying individual human reasoning styles. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5038–5076.

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.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2302–2310.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *SciPy*, 2015:18–24.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. 2025. Missrag: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3215–3224.

Jose C Principe, Dongxin Xu, Qun Zhao, and John W Fisher Iii. 2000. Learning from examples with information theoretic criteria. *Journal of VLSI signal processing systems for signal, image and video technology*, 26(1):61–77.

Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. 2025. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 742–754.

Kyeongha Rho, Hyeongkeun Lee, Valentio Iverson, and Joon Son Chung. 2025. Lavcap: Llm-based audio-visual captioning using optimal transport. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jianwen Sun, Yukang Feng, Yifan Chang, Chuanhao Li, Zizhen Li, Jiaxin Ai, Fanrui Zhang, Yu Dai, and Kaipeng Zhang. 2025a. Closing the expression gap in llm instructions via socratic questioning. *arXiv preprint arXiv:2510.27410*.

Jianwen Sun, Yukang Feng, Chuanhao Li, Fanrui Zhang, Zizhen Li, Jiaxin Ai, Sizhuo Zhou, Yu Dai, Shenglin Zhang, and Kaipeng Zhang. 2025b. Armor: Empowering multimodal understanding model with interleaved multimodal generation capability. *arXiv preprint arXiv:2503.06542*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

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

Haining Wang, Kang He, Bobo Li, Lei Chen, Fei Li, Xu Han, Chong Teng, and Donghong Ji. 2024. Refining and synthesis: A simple yet effective data augmentation framework for cross-domain aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10318–10329.

Hu Wang, Jianpeng Zhang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2022b. Uncertainty-aware multi-modal learning via cross-modal random network prediction. In *European Conference on Computer Vision*, pages 200–217.

Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. 2024. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pages 71–86.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515.

Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4730–4738.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

Wenzhe Yin, Zehao Xiao, Pan Zhou, Shujian Yu, Jiayi Shen, Jan-Jakob Sonke, and Efstratios Gavves. 2025. Distributional vision-language alignment by cauchy-schwarz divergence. *arXiv preprint arXiv:2502.17028*.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10790–10797.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4400–4407.

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

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. *Advances in Neural Information Processing Systems*, 37:55943–55974.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023a. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023b. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767.

Miao Zhou, Lina Yang, Thomas Wu, Dongnan Yang, and Xinru Zhang. 2025. Dual-path dynamic fusion with learnable query for multimodal sentiment analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11355–11365.

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

A Experimental Setup

In this section, we provide a comprehensive description of the experimental setup. Specifically, we elaborate on the dataset characteristics, introduce the baseline models used for comparison, and specify the implementation details, along with the experimental designs for our in-depth analysis.

A.1 Datasets

We evaluate EASE on three multimodal sentiment analysis benchmarks that differ in language, scale, and annotation granularity. **MOSI** (Zadeh et al., 2016) contains 2,199 opinion video clips from YouTube with sentiment scores in [-3, +3], serving as a compact benchmark for robustness under limited data. **MOSEI** (Bagher Zadeh et al., 2018) extends MOSI to 22,856 segments across 250 topics with richer speaker diversity, enabling evaluation of generalization under missing-modality settings. **SIMS** (Yu et al., 2020) is a Chinese dataset of 2,281 clips with sentiment scores in [-1, +1], introducing cross-lingual and cross-cultural challenges.

A.2 Evaluation Metrics

To ensure statistical reliability and a rigorous comparison with existing benchmarks, we adhere to standard MSA protocols. All reported results represent the average performance initialized with different random seeds.

Regression and Continuous Evaluation. We employ Mean Absolute Error (MAE) and Pearson Correlation (Corr) to evaluate the model's capability in predicting continuous sentiment intensity. MAE quantifies the average magnitude of errors between predicted and ground-truth scores, where lower values indicate superior performance. Conversely, Corr measures the linear relationship between predictions and targets, with higher values reflecting a better capture of relative sentiment trends.

Classification Standards. Regarding classification, we adopt metrics tailored to the granularity of each dataset. For the English datasets (MOSI and MOSEI), we report 7-class accuracy (Acc-7) to assess performance across the full sentiment spectrum from -3 to $+3$, as well as 5-class accuracy (Acc-5). For binary classification, following established conventions (Yu et al., 2021; Zhang et al., 2024), we report Binary Accuracy (Acc-2) and F1 scores in two distinct configurations, denoted in tables as *"Non-negative / Positive"*. The value on the left corresponds to a standard classification separating Negative (< 0) from Non-negative (≥ 0) samples, thereby including neutral instances in the positive class. The value on the right represents a stricter classification separating Negative (< 0) from Positive (> 0) samples, where neutral (zero-valued) instances are explicitly excluded from the evaluation. For the Chinese dataset (SIMS), in addition to the binary metrics described above, we further include 3-class accuracy (Acc-3) for Negative/Neutral/Positive classification and 5-class accuracy (Acc-5) for fine-grained analysis.

A.3 Implementation Details

To ensure a fair comparison, we align our evaluation settings with the robust protocols in prior work LNLN (Zhang et al., 2024). We assess model performance under two distinct missingness paradigms: random missingness (*Intra-Modality*) and fixed missingness (*Inter-Modality*).

We systematically evaluate the resilience of our proposed EASE by varying the missing rate MR from 0 to 0.9 with a step size of 0.1. This setting simulates continuous varying degrees of data incompleteness. We further evaluate performance under specific modality-missing scenarios to test robustness in extreme cases. Specifically, we test on all possible incomplete subsets, including $\{l\}$, $\{a\}$, $\{v\}$, $\{l, a\}$, $\{l, v\}$, $\{a, v\}$.

We implement EASE using the PyTorch framework. All models are trained for 200 epochs with a batch size of 64 using the Adam optimizer, a fixed learning rate of e^{-5}, and a single NVIDIA RTX 4090 GPU. The EASE framework is trained in a stage-wise manner, and all hyperparameters are tuned on the validation set. Unless otherwise specified, we set $\lambda_1 = 0.5$, $\lambda_2 = 0.2$, $\lambda_3 = 0.5$, and $\beta_1 = 0.02$, which yields stable convergence across all datasets.

A.4 Feature Extraction

To guarantee rigorous benchmarking against SOTA approaches, we align our experimental setup with the MSA evaluation protocol (Wang et al., 2023; Li et al., 2025a), utilizing their standardized pre-computed features. The extraction details for each modality are outlined as follows:

Text Modality. Linguistic features are obtained using the `BERT-base-uncased` model (Devlin et al., 2019) for the English datasets (MOSI and

MOSEI) and `BERT-base-chinese` for SIMS. We maintain a unified feature dimension of 768 across all tasks, with input sequences truncated to lengths of 50 for MOSI/MOSEI and 39 for SIMS.

Audio Modality. Acoustic signals in MOSI and MOSEI are processed via COVAREP (Degottex et al., 2014) to extract 5 and 74 low-level acoustic features, including pitch and MFCCs. The input sequence lengths are 375 and 500, respectively. For SIMS, we employ Librosa(McFee et al., 2015) to extract 33-dimensional features with a sequence length of 400.

Visual Modality. For the English datasets, Facet is applied to encode facial behaviors (*e.g.*, Action Units), yielding feature dimensions of 20 (MOSI) and 35 (MOSEI), both aligned to a length of 500. Conversely, visual features for SIMS are extracted using OpenFace 2.0 (Baltrušaitis et al., 2016), producing 709-dimensional vectors with a sequence length of 55.

A.5 Baselines

To verify the effectiveness of our proposed method, we conduct a comprehensive comparison against a diverse set of advanced baselines.

Complete-Modality Methods. These approaches focus on sophisticated fusion mechanisms to maximize multimodal information interaction:

MISA (Hazarika et al., 2020): projects features into modality-invariant and -specific subspaces to learn a holistic view of the data.

SelfMM (Yu et al., 2021): generates unimodal labels for joint multi-task training to enhance the capture of modality-specific cues.

MMIM (Han et al., 2021): maximizes mutual information in unimodal pairs and fusion results to preserve task-relevant information.

CENet (Wang et al., 2022a): embeds non-verbal cues into pre-trained language models via feature transformation to enhance text representations.

TETFN (Wang et al., 2023): utilizes text-guided attention and cross-modal mappings to capture sentiment cues while retaining modality-specific predictions.

ALMT (Zhang et al., 2023b): incorporates a hyper-modality learning module to suppress irrelevant noise under the guidance of language features.

Missing-Modality Methods. To evaluate performance under realistic scenarios with data loss, we compare against robust frameworks that employ reconstruction or adaptation strategies:

TFR-Net (Yuan et al., 2021): generates missing features via a Transformer-based reconstruction network supervised by Smooth L1 Loss.

LNLN (Zhang et al., 2024): employs a language-dominated framework with dominant-modality correction to handle noisy or missing data.

P-RMF (Zhu et al., 2025): maps unimodal data to Gaussian latent spaces to learn stable representations via proxy-driven fusion.

TF-Mamba (Li et al., 2025a): utilizes text-aware enhancement and text-guided Mamba modules to efficiently model long sequences with missing modalities.

A.6 Elastic Alignment Experiments

Experimental Setup. First, we utilize the pre-trained EASE model to extract the completed features $\hat{\mathbf{z}}$ and their corresponding ground truth representations $\mathbf{z}^{gt}$ for all test samples. To systematically observe the model's behavior across varying difficulty levels, we quantify the reconstruction difficulty of each sample using the Mean Squared Error (MSE: $\|\mathbf{z}^{gt} - \hat{\mathbf{z}}\|_2^2$). Accordingly, all test samples are sorted by this metric and mapped to a percentile rank range of $[0, 100]$.

We then contrast two distinct penalty paradigms to visualize the impact of uncertainty calibration. ❶: **Rigid Baseline** simulates traditional deterministic alignment methods (*e.g.*, standard MSE or KL loss), where the alignment penalty is strictly proportional to the raw reconstruction error (*i.e.*, Penalty $\propto$ Error). ❷: **EASE (Elastic)** represents the effective penalty enforced by EASE, calculated as Error$/\psi(u)$, where $\psi(u)$ is the uncertainty-adaptive bandwidth derived in Eq. (6).

Note: The Rationality of Low Penalty in Language. The language modality exhibits an extreme case of global relaxation (penalty ≈ 0.4). This reflects the *one-to-many ambiguity* inherent in the inverse generation task. For instance, a visual *smile* may correspond to diverse textual descriptions like "*happy*", "*joyful*", or "*great*". Rigid alignment forces the model to strictly fit one specific token, leading to overfitting. Conversely, EASE identifies this high aleatoric uncertainty and reduces the alignment cost, encouraging the model to preserve core sentiment semantics.

B Theoretical Justification and Derivation

B.1 Theoretical Motivation and Derivation for Gaussian Modeling

In the UCMI module, we model the conditional distribution of missing features $q(z_B|\tilde{\mathbf{x}})$ as a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_B, \boldsymbol{\sigma}_B^2)$. The validity of this assumption is grounded in two key aspects of our framework design:

Semantic Continuity and Unimodality. In the high-dimensional latent space constructed by deep neural networks, the semantic representation of a specific utterance typically forms a continuous and unimodal cluster. The multivariate Gaussian distribution naturally captures this local structure, where the mean $\boldsymbol{\mu}_B$ represents the semantic centroid and the variance $\boldsymbol{\sigma}_B^2$ models the spherical spread of potential variations caused by missing modality ambiguity.

Analytical Tractability for Elastic Alignment. A critical contribution of EASE is the Uncertainty-Calibrated Semantic Alignment (UCSA), which relies on the Cauchy-Schwarz (CS) divergence. The Gaussian assumption allows us to derive a **closed-form solution** for the CS divergence (as detailed in Appendix B.2), explicitly decomposing the objective into a *semantic centrality term* and a *distribution shape term*. Without this parametric assumption, calculating the divergence would require computationally expensive Monte Carlo sampling, hindering efficient end-to-end training.

Derivation of the Optimization Objective. Based on this Gaussian assumption, we derive the optimization objective for the UCMI module via Maximum Likelihood Estimation (MLE). Let z^{gt} be the ground truth feature vector. We aim to maximize the likelihood of the observed ground truth under the predicted distribution:

$$\mathcal{L}_{MLE} = \sum_{i=1}^{N} \log p(z_i^{gt}|\tilde{\mathbf{x}}_i; \phi) \tag{14}$$

Based on the Gaussian assumption with a diagonal covariance matrix $\text{diag}(\boldsymbol{\sigma}^2)$, the dimensions of the feature vector are statistically independent. Substituting the multivariate Gaussian Probability Density Function (PDF) with diagonal covariance:

$$\begin{aligned}\log p(z^{gt}|\tilde{\mathbf{x}}) &= \sum_{d=1}^{D} \log \mathcal{N}(z_d^{gt}; \mu_d, \sigma_d^2) \\ &= \sum_{d=1}^{D} \log \left(\frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left(-\frac{(z_d^{gt} - \mu_d)^2}{2\sigma_d^2} \right) \right) \\ &= \sum_{d=1}^{D} \left(-\frac{1}{2}\log(2\pi\sigma_d^2) - \frac{(z_d^{gt} - \mu_d)^2}{2\sigma_d^2} \right)\end{aligned} \tag{15}$$

Maximizing the log-likelihood is equivalent to minimizing the Negative Log-Likelihood (NLL):

$$\begin{aligned}\mathcal{L}_{\text{ucmi}} &= -\log p(z^{gt}|\tilde{\mathbf{x}}) \\ &= \frac{1}{2}\sum_{d=1}^{D} \left(\log(2\pi\sigma_d^2) + \frac{(z_d^{gt} - \mu_d)^2}{\sigma_d^2} \right)\end{aligned} \tag{16}$$

This derivation proves that our loss function is theoretically consistent with maximizing the probability of the true features given the partial input, while simultaneously learning the uncertainty σ^2 as a heteroscedastic noise estimator.

B.2 Derivation of Closed-Form CS Divergence

We derive the analytic expression for the Cauchy-Schwarz (CS) divergence between the completed feature distribution $\mathcal{Q} = \mathcal{N}(\hat{\boldsymbol{z}}, \hat{\boldsymbol{\Sigma}})$ and the semantic anchor $\mathcal{P} = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

The CS divergence is defined as:

$$D_{CS}(\mathcal{Q}\|\mathcal{P}) = -\log \left(\frac{\left(\int \mathcal{Q}(\mathbf{x})\mathcal{P}(\mathbf{x})d\mathbf{x}\right)^2}{\int \mathcal{Q}(\mathbf{x})^2 d\mathbf{x} \int \mathcal{P}(\mathbf{x})^2 d\mathbf{x}} \right). \tag{17}$$

The derivation relies on the product identity of two Gaussians. For $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, the integral of their product is:

$$\int \mathcal{N}_1(\mathbf{x})\mathcal{N}_2(\mathbf{x})d\mathbf{x} = \mathcal{N}(\boldsymbol{\mu}_1; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2). \tag{18}$$

Step A: Cross-term. Applying the Gaussian product identity with $\boldsymbol{\mu}_1 = \hat{\boldsymbol{z}}, \boldsymbol{\Sigma}_1 = \hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_k$, we have

$$\int \mathcal{Q}\mathcal{P}d\mathbf{x} = V_{\mathcal{Q}\mathcal{P}} = \mathcal{N}(\hat{\boldsymbol{z}}; \boldsymbol{\mu}_k, \hat{\boldsymbol{\Sigma}} + \boldsymbol{\Sigma}_k). \tag{19}$$

Taking the logarithm and multiplying by -2 yields

$$\begin{aligned}-2\log V_{\mathcal{Q}\mathcal{P}} = &\log\left((2\pi)^D|\hat{\boldsymbol{\Sigma}} + \boldsymbol{\Sigma}_k|\right) \\ &+ (\hat{\boldsymbol{z}} - \boldsymbol{\mu}_k)^\top (\hat{\boldsymbol{\Sigma}} + \boldsymbol{\Sigma}_k)^{-1} (\hat{\boldsymbol{z}} - \boldsymbol{\mu}_k).\end{aligned} \tag{20}$$

Step B: Self-terms. For self-term of $\mathcal{Q}$, we have

$$\begin{aligned}\int \mathcal{Q}^2 d\mathbf{x} &= \int \mathcal{N}(\mathbf{x}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{\Sigma}})\, \mathcal{N}(\mathbf{x}; \hat{\boldsymbol{z}}, \hat{\boldsymbol{\Sigma}})\, d\mathbf{x} \\ &= \mathcal{N}(\hat{\boldsymbol{z}}; \hat{\boldsymbol{z}}, 2\hat{\boldsymbol{\Sigma}}) = \frac{1}{\sqrt{(2\pi)^D |2\hat{\boldsymbol{\Sigma}}|}},\end{aligned} \tag{21}$$

and therefore

$$\log \int \mathcal{Q}^2 d\mathbf{x} = -\frac{1}{2} \log \left((2\pi)^D |2\hat{\mathbf{\Sigma}}| \right). \tag{22}$$

Similarly, for $\mathcal{P}$ we obtain

$$\log \int \mathcal{P}^2 d\mathbf{x} = -\frac{1}{2} \log \left((2\pi)^D |2\mathbf{\Sigma}_k| \right). \tag{23}$$

Step C: Final Closed Form. Substituting the three terms back into the definition of D_{CS} and canceling the common $(2\pi)^D$ factor, we arrive at

$$D_{CS}(\mathcal{Q}\|\mathcal{P}) = (\hat{\boldsymbol{z}} - \boldsymbol{\mu}_k)^\top (\hat{\mathbf{\Sigma}} + \mathbf{\Sigma}_k)^{-1} (\hat{\boldsymbol{z}} - \boldsymbol{\mu}_k) + \log \frac{|\hat{\mathbf{\Sigma}} + \mathbf{\Sigma}_k|}{\sqrt{|2\hat{\mathbf{\Sigma}}| \cdot |2\mathbf{\Sigma}_k|}}. \tag{24}$$

We denote this quantity as the anchor loss:

$$\mathcal{L}_{\text{anchor}} = \underbrace{(\hat{\boldsymbol{z}} - \boldsymbol{\mu}_k)^\top (\hat{\mathbf{\Sigma}} + \mathbf{\Sigma}_k)^{-1} (\hat{\boldsymbol{z}} - \boldsymbol{\mu}_k)}_{\text{Mahalanobis distance term}} + \underbrace{\log \frac{|\hat{\mathbf{\Sigma}} + \mathbf{\Sigma}_k|}{\sqrt{|2\hat{\mathbf{\Sigma}}| \cdot |2\mathbf{\Sigma}_k|}}}_{\text{Log-determinant term}}. \tag{25}$$

The final closed form decomposes into a *center* term and a *shape* term:

Mahalanobis distance term. The first term measures the distance between the completed feature mean $\hat{\boldsymbol{z}}$ and the anchor mean $\boldsymbol{\mu}_k$ under the combined covariance $(\hat{\mathbf{\Sigma}} + \mathbf{\Sigma}_k)$. Because this covariance appears in the denominator, large uncertainty (large variance) automatically attenuates the contribution of this distance, which is consistent with the design philosophy of UADA: uncertain completions are not forced to match the anchor center too aggressively.

Log-determinant term. The second term compares the "volume" (or shape) of the two Gaussian distributions through their determinants. It encourages the covariance of the completed features $\hat{\mathbf{\Sigma}}$ to be compatible with the anchor covariance $\mathbf{\Sigma}_k$, preventing the hallucinated distribution from being overly collapsed or overly dispersed.

Taken together, the two terms implement a joint alignment of *distribution centers* and *distribution shapes*, providing a principled semantic-anchoring mechanism for the hallucinated features.

Implementation Details of Distribution Parameters. To ensure numerical stability and computational efficiency, we employ diagonal covariance matrices for both the predicted distribution $\mathcal{Q}$ and the semantic anchor $\mathcal{P}$.

Predicted Covariance $\hat{\mathbf{\Sigma}}$. The predicted covariance matrix is derived from the uncertainty head of the UCMI module. Specifically, the network predicts a variance vector $\hat{\boldsymbol{\sigma}}^2 \in \mathbb{R}^D$, and we enforce positivity using the Softplus function with a small stability constant $\epsilon = 1e^{-6}$:

$$\hat{\mathbf{\Sigma}} = \text{diag}(\hat{\boldsymbol{\sigma}}^2) \tag{26}$$

$$\hat{\boldsymbol{\sigma}}^2 = \text{Softplus}(\mathcal{W}_\sigma h + b_\sigma) + \epsilon \tag{27}$$

Here, h is the latent representation and $\{W_\sigma, b_\sigma\}$ are learnable parameters.

Semantic Anchor Covariance $\mathbf{\Sigma}_k$. The semantic anchors represent the global statistical characteristics of each sentiment class. We compute $\mathbf{\Sigma}_k$ as the empirical diagonal covariance of the features belonging to class k in the training set. Let $\mathcal{Z}_k = \{z_i \mid y_i = k\}$ be the set of features from complete modality samples with label k. The anchor parameters are updated as:

$$\boldsymbol{\mu}_k = \frac{1}{|\mathcal{Z}_k|} \sum_{z \in \mathcal{Z}_k} z \tag{28}$$

$$\mathbf{\Sigma}_k = \text{diag}\left(\frac{1}{|\mathcal{Z}_k|} \sum_{z \in \mathcal{Z}_k} (z - \boldsymbol{\mu}_k)^2 \right) + \epsilon \mathbf{I}. \tag{29}$$

In practice, these statistics are initialized using the pre-trained feature extractor and can be updated via a moving average strategy during training to maintain semantic stability.

C More Experimental Results

C.1 Intra-Modality Robustness Results

Tables 7, 8 and 9 present a granular comparison of intra-modality robustness on the MOSI, MOSEI, and SIMS datasets. We observe a distinct performance divergence relative to the noise intensity. In low-noise regimes (missing rate $r < 0.4$), strong baselines such as Self-MM and LNLN maintain competitive performance, effectively exploiting the fine-grained semantics present in nearly complete data. However, as the missing rate escalates ($r \geq 0.5$), EASE demonstrates significant superiority, consistently achieving optimal results in Correlation and MAE metrics. This confirms that EASE effectively mitigates high-intensity noise interference via its uncertainty-calibrated mechanism,

Condition	MOSI						MOSEI					
	Acc-2	F1	Acc-5	Acc-7	MAE	Corr	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
w/o $\{v\}$	81.69 / 83.70	81.95 / 83.87	49.02	44.23	0.752	0.775	83.44 / 84.79	83.71 / 85.06	52.66	50.15	0.548	0.739
w/o $\{a\}$	81.50 / 83.44	81.76 / 83.60	49.37	44.19	0.740	0.778	83.32 / 84.65	83.40 / 84.98	52.70	50.27	0.549	0.742
w/o $\{l\}$	59.89 / 60.85	66.74 / 68.61	25.37	23.19	1.314	0.184	71.02 / 67.57	82.82 / 77.73	41.82	39.23	0.810	0.252

Table 6: Single-modality ablations of EASE on the MOSI and MOSEI datasets. Refer to Sec. 4.4.

whereas rigid baselines tend to overfit corrupted residuals.

Discussion on Trade-offs. Generally, models trained with heavy noise augmentation struggle to maintain SOTA precision on clean data due to distribution shifts: a phenomenon known as the Robustness-Accuracy Trade-off. While EASE alleviates this issue through uncertainty estimate and elastic alignment, achieving a perfect equilibrium between robustness under extreme noise and high fidelity on clean data remains a pivotal direction for future research.

C.2 Inter-Modality Robustness Results

Table 10 presents the generalization performance on MOSI and MOSEI under various *inter-modality missing* settings. Overall, methods exhibit noticeably different robustness behaviors depending on which modality is absent. When the language modality is preserved (*e.g.*, $\{l\}$, $\{l, a\}$, $\{l, v\}$), most baselines remain competitive, reflecting the dominant role of textual cues in sentiment prediction. In contrast, removing language-related information (*e.g.*, $\{a\}$, $\{v\}$, $\{a, v\}$) leads to substantial performance degradation for many methods. Across these challenging settings, EASE consistently achieves stronger Acc-5/Acc-7 performance and lower MAE, indicating improved resilience to severe cross-modal information loss. These results suggest that uncertainty-aware elastic alignment enables EASE to better suppress unreliable modality contributions and maintain stable predictions under diverse missing-modality scenarios.

C.3 Ablation Study on Modality

To evaluate the individual contribution of each modality to the final sentiment prediction, we conduct single-modality ablation studies on the MOSI and MOSEI datasets. Table 6 results reveal a distinct hierarchy in modal importance.

First, we observe the *dominance of linguistics*. Scenarios retaining text (*e.g.*, w/o v) maintain superior performance, with MOSEI achieving a high Acc-7 of 50.15%, confirming language as the primary semantic backbone. In contrast, the *complementary nature of non-verbal cues* is evident when removing text (w/o l): Acc-7 on MOSI plummets to 23.19% and MAE nearly doubles to 1.314. This indicates that while acoustic and visual signals are insufficient for fine-grained reasoning alone, EASE leverages them as critical supplements to resolve ambiguities in the foundational linguistic representations.

C.4 More Visualization Results

We further present elastic alignment and predictive consistency analyses on the MOSI dataset. As shown in Fig. 8, EASE adaptively suppresses excessive alignment penalties in regions with high reconstruction uncertainty, especially for visual and acoustic modalities, preventing noisy or ambiguous samples from dominating optimization. This elastic behavior leads to a smoother and more stable alignment profile compared to rigid baselines.

In Fig. 9, consistency analysis shows that incorporating MIPC substantially reduces distribution shift between full-modality and missing-modality predictions and enforces tighter instance-level agreement. Together, these results demonstrate that EASE achieves robust sentiment modeling by jointly balancing uncertainty-aware alignment and cross-view predictive consistency under modality missingness.

Method	Acc-2↑	F1↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑	Method	Acc-2↑	F1↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
Random Missing Rate $r = 0$							Random Missing Rate $r = 0.5$						
MISA	81.24 / 82.78	81.23 / 82.83	48.30	43.05	0.771	0.777	MISA	69.34 / 70.53	69.20 / 70.50	30.61	28.14	1.124	0.519
Self-MM	**83.24 / 85.22**	**83.26 / 85.19**	**52.38**	42.81	**0.720**	**0.790**	Self-MM	67.54 / 67.43	66.81 / 64.27	31.39	26.97	1.129	0.503
MMIM	81.97 / 83.43	81.94 / 83.43	49.85	**45.92**	0.744	0.778	MMIM	66.52 / 68.09	64.59 / 66.15	29.89	28.23	1.128	0.501
TFR-Net	81.68 / 83.64	81.61 / 83.57	47.91	40.82	0.805	0.760	TFR-Net	63.02 / 64.83	56.64 / 58.04	30.71	25.85	1.270	0.443
CENET	81.49 / 83.08	81.48 / 83.06	50.39	43.20	0.748	0.785	CENET	66.08 / 72.46	63.50 / 71.10	30.90	28.33	1.130	0.496
ALMT	82.75 / 84.91	82.94 / 85.01	48.49	42.37	0.752	0.768	ALMT	65.94 / 68.24	68.54 / 69.74	31.25	28.42	1.138	0.485
LNLN	81.24 / 84.25	81.79 / 84.61	49.76	44.56	0.751	0.778	LNLN	71.86 / 73.37	72.30 / 73.70	38.39	33.92	1.059	0.536
TF-Mamba	81.63 / 83.69	81.58 / 83.71	50.58	44.31	0.762	0.774	TF-Mamba	74.20 / 75.00	74.27 / 75.15	37.46	33.67	1.044	0.557
P-RMF	82.65 / 84.15	82.69 / 84.37	48.83	44.31	0.726	0.782	P-RMF	71.28 / 73.02	71.66 / 73.33	37.90	33.67	1.077	0.523
EASE	82.81 / 84.87	82.85 / 84.90	50.20	46.73	0.748	0.779	**EASE**	**74.43 / 75.49**	**74.40 / 75.57**	**38.44**	**34.64**	**0.998**	**0.578**
Random Missing Rate $r = 0.1$							Random Missing Rate $r = 0.6$						
MISA	76.34 / 77.54	76.30 / 77.58	41.55	36.25	0.939	0.654	MISA	65.84 / 66.97	65.69 / 66.94	27.12	24.68	1.200	0.441
Self-MM	76.48 / 78.15	76.51 / 77.76	43.98	36.64	0.901	0.660	Self-MM	63.36 / 63.47	62.07 / 58.94	27.31	24.34	1.209	0.425
MMIM	74.54 / 76.42	74.22 / 76.12	42.66	39.07	0.918	0.651	MMIM	62.49 / 63.67	59.48 / 60.87	27.11	25.41	1.208	0.418
TFR-Net	73.52 / 74.70	72.70 / 73.57	40.13	34.70	0.987	0.622	TFR-Net	59.47 / 61.64	50.53 / 52.44	28.33	24.05	1.371	0.363
CENET	74.64 / 77.49	74.28 / 77.35	42.32	38.00	0.916	0.654	CENET	61.47 / 67.58	57.86 / 64.87	26.53	24.54	1.215	0.415
ALMT	75.70 / 77.64	76.24 / 77.94	40.33	35.33	0.927	0.645	ALMT	62.15 / 64.53	65.87 / 66.81	27.36	25.41	1.214	0.407
LNLN	78.43 / 81.20	79.04 / 81.62	47.91	42.37	0.820	0.724	LNLN	67.69 / 69.00	67.99 / 69.19	34.35	30.37	1.147	0.458
TF-Mamba	80.03 / 81.86	79.97 / 81.87	**48.40**	**42.86**	**0.824**	**0.732**	TF-Mamba	68.37 / 68.60	68.45 / 68.79	33.53	30.76	1.127	0.487
P-RMF	**81.34 / 82.62**	**81.35 / 82.89**	47.52	42.13	0.800	0.730	P-RMF	67.35 / 68.75	67.64 / 68.64	33.24	29.30	1.147	0.432
EASE	80.87 / 81.98	80.74 / 82.01	48.63	44.30	0.816	0.721	**EASE**	**71.70 / 72.07**	**71.75 / 72.18**	**35.23**	**32.19**	**1.065**	**0.526**
Random Missing Rate $r = 0.2$							Random Missing Rate $r = 0.7$						
MISA	74.54 / 75.76	74.51 / 75.82	38.97	34.60	0.989	0.618	MISA	63.89 / 65.09	63.74 / 65.07	23.27	21.14	1.257	0.381
Self-MM	74.98 / 76.37	74.94 / 75.68	40.67	34.89	0.967	0.614	Self-MM	61.46 / 61.74	58.97 / 55.11	23.81	20.70	1.271	0.339
MMIM	71.91 / 74.08	71.28 / 73.47	40.43	36.83	0.974	0.612	MMIM	59.18 / 61.23	54.36 / 57.15	24.00	22.35	1.267	0.342
TFR-Net	71.28 / 72.36	69.58 / 70.12	38.34	32.55	1.065	0.572	TFR-Net	57.34 / 59.91	45.48 / 48.41	26.92	23.71	1.454	0.276
CENET	72.01 / 76.83	71.30 / 76.56	38.97	34.74	0.983	0.605	CENET	59.43 / 63.82	54.22 / 53.79	23.57	22.35	1.269	0.335
ALMT	72.94 / 75.15	73.66 / 75.51	37.17	33.04	0.992	0.596	ALMT	59.67 / 61.84	65.19 / 65.30	24.97	23.71	1.266	0.336
LNLN	76.87 / 79.22	77.34 / 79.53	**45.14**	39.74	0.891	0.668	LNLN	65.01 / 65.95	65.14 / 65.95	31.19	27.79	1.219	0.383
TF-Mamba	**79.15 / 80.49**	**79.17** / 80.56	44.75	39.21	0.879	**0.693**	TF-Mamba	66.91 / 67.23	66.98 / 67.41	29.30	27.26	1.196	0.411
P-RMF	78.13 / 79.57	78.11 / **80.97**	44.75	40.38	**0.853**	0.668	P-RMF	65.16 / 66.16	65.33 / 64.69	32.94	27.84	1.229	0.383
EASE	79.11 / 80.25	79.19 / 80.32	45.87	**42.41**	0.857	0.685	**EASE**	**68.19 / 69.36**	**68.17 / 69.30**	**33.51**	**28.85**	**1.153**	**0.487**
Random Missing Rate $r = 0.3$							Random Missing Rate $r = 0.8$						
MISA	74.54 / 75.76	74.51 / 75.82	38.97	34.60	0.989	0.618	MISA	62.24 / 63.56	61.67 / 63.16	20.99	19.92	1.311	0.321
Self-MM	74.98 / 76.37	74.94 / 75.68	40.67	34.89	0.967	0.614	Self-MM	58.26 / 59.55	53.56 / 49.98	22.11	19.29	1.313	0.282
MMIM	71.91 / 74.08	71.28 / 73.47	40.43	36.83	0.974	0.612	MMIM	55.30 / 58.33	47.89 / 52.46	21.77	20.26	1.312	0.287
TFR-Net	71.28 / 72.36	69.58 / 70.12	38.34	32.55	1.065	0.572	TFR-Net	55.98 / 58.49	41.88 / 44.70	27.70	23.23	1.497	0.155
CENET	72.01 / 76.83	71.30 / 76.56	38.97	34.74	0.983	0.605	CENET	57.53 / 60.93	50.80 / 54.68	21.67	21.14	1.314	0.274
ALMT	72.94 / 75.15	73.66 / 75.51	37.17	33.04	0.992	0.596	ALMT	58.31 / 60.37	**66.14** / 65.45	23.66	23.13	1.310	0.273
LNLN	75.46 / 77.29	75.68 / 77.56	42.81	38.00	0.953	0.617	LNLN	62.10 / 62.75	62.03 / 62.56	28.23	26.34	1.283	0.314
TF-Mamba	76.53 / 77.74	76.56 / 77.85	42.27	37.76	**0.932**	**0.645**	TF-Mamba	63.12 / 63.57	63.20 / 63.77	26.38	24.93	1.258	0.353
P-RMF	75.80 / 76.83	75.82 / **79.27**	42.71	**39.21**	0.922	0.621	P-RMF	61.08 / 62.04	61.22 / 60.76	29.74	25.97	1.275	0.316
EASE	**76.89 / 77.94**	**76.92** / 77.98	**43.69**	39.12	**0.894**	0.630	**EASE**	**64.63 / 65.51**	64.89 / **65.73**	**30.28**	**27.18**	**1.216**	**0.405**
Random Missing Rate $r = 0.4$							Random Missing Rate $r = 0.9$						
MISA	72.59 / 73.88	72.49 / 73.88	35.37	32.65	1.041	0.585	MISA	58.21 / 58.64	56.19 / 56.84	18.41	17.78	1.369	0.226
Self-MM	71.96 / 73.17	71.75 / 71.74	36.30	31.20	1.027	0.579	Self-MM	55.25 / 58.59	47.46 / 46.16	19.78	18.32	1.353	0.197
MMIM	68.90 / 70.84	67.80 / 69.69	35.76	33.38	1.034	0.576	MMIM	51.65 / 55.29	40.89 / 47.33	19.53	18.95	1.357	0.186
TFR-Net	67.74 / 68.75	64.41 / 64.71	35.76	30.17	1.142	0.537	TFR-Net	55.44 / 57.93	40.18 / 43.01	25.12	21.67	1.534	0.155
CENET	71.53 / 73.38	70.26 / 72.75	36.15	32.26	1.031	0.574	CENET	54.76 / 58.99	46.58 / 50.01	19.10	19.15	1.357	0.181
ALMT	71.14 / 73.12	72.47 / 73.85	35.03	31.44	1.045	0.560	ALMT	56.66 / 57.32	**67.82 / 64.92**	20.50	20.31	1.349	0.205
LNLN	74.25 / 76.01	74.67 / 76.31	41.11	36.49	0.987	0.594	LNLN	56.51 / 56.50	56.47 / 56.32	23.86	22.98	1.349	0.202
TF-Mamba	75.22 / 76.07	75.25 / 76.16	40.23	35.86	0.961	**0.617**	TF-Mamba	60.20 / 60.37	60.30 / 60.59	24.49	22.89	1.363	0.215
P-RMF	73.76 / 75.46	74.09 / 77.71	40.67	35.59	1.001	0.584	P-RMF	58.75 / 59.45	59.01 / 56.66	26.68	23.49	1.346	0.212
EASE	**75.45 / 76.36**	**75.48 / 76.40**	**41.12**	**36.89**	**0.924**	0.604	**EASE**	**62.06 / 63.24**	62.35 / 63.52	**27.74**	**25.16**	**1.283**	**0.312**

Table 7: Details of robust comparison on MOSI under different intra-modality missing rates. Refer to Sec. 4.3 for details.

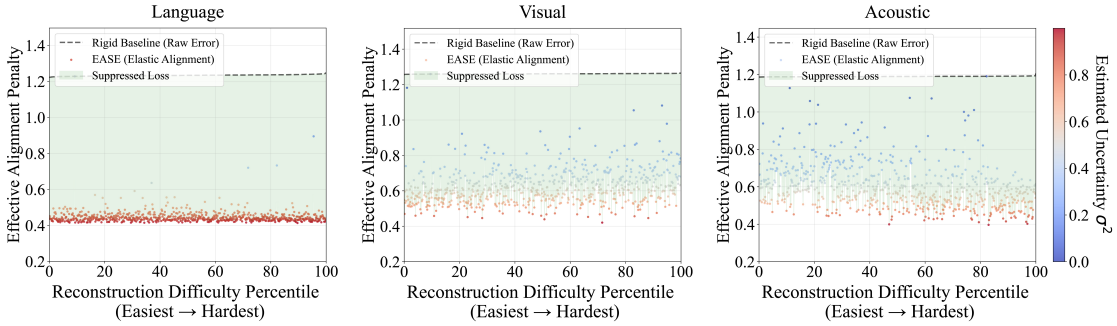
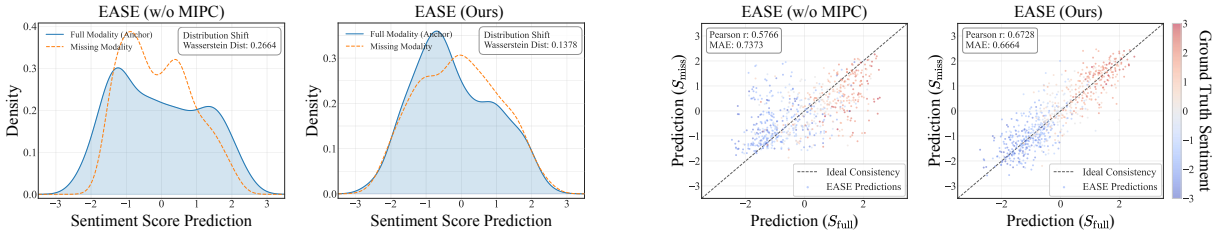


Figure 8: Elastic Alignment Visualization on MOSI. Best viewed in color. Refer to Sec. 4.5.

Method	Acc-2↑	F1↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑	Method	Acc-2↑	F1↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
Random Missing Rate $r = 0$							Random Missing Rate $r = 0.5$						
MISA	84.10 / 85.28	83.75 / 85.10	53.85	51.79	0.552	0.759	MISA	73.21 / 67.38	64.14 / 58.38	36.05	38.12	0.834	0.492
Self-MM	**84.68** / 85.34	**84.66** / 85.11	55.72	53.89	**0.531**	0.764	Self-MM	75.81 / 71.97	70.38 / 67.40	43.14	42.70	0.733	0.477
MMIM	81.65 / 83.53	81.41 / 83.39	53.04	50.76	0.576	0.724	MMIM	74.45 / 71.75	67.96 / 67.70	39.21	38.68	0.775	0.470
TFR-Net	84.65 / 84.96	84.34 / 84.71	47.91	53.71	0.550	0.745	TFR-Net	75.69 / 71.53	70.07 / 66.88	30.71	45.00	0.730	0.471
CENET	82.30 / 85.49	82.60 / 85.41	**56.12**	**54.39**	**0.531**	**0.770**	CENET	77.16 / 73.33	74.14 / 69.80	45.52	45.12	0.720	0.515
ALMT	83.99 / **85.62**	84.53 / **85.69**	53.89	52.18	0.542	0.752	ALMT	77.48 / 77.40	77.80 / 77.73	38.34	37.82	0.683	0.461
LNLN	83.61 / 84.14	84.02 / 84.53	51.94	50.66	0.572	0.735	LNLN	78.10 / 76.44	79.30 / 77.23	45.59	44.90	0.710	0.529
TF-Mamba	82.89 / 83.82	82.92 / 83.71	53.83	52.26	0.556	0.748	TF-Mamba	77.94 / 78.59	77.78 / 78.34	**46.60**	45.68	0.676	0.583
P-RMF	83.62 / 85.20	83.68 / 85.48	52.09	49.77	0.539	0.767	P-RMF	78.64 / 78.62	79.74 / 79.49	44.90	43.94	0.666	0.601
EASE	83.91 / 85.55	83.97 / 85.62	54.82	52.48	0.543	0.765	**EASE**	**79.49 / 79.12**	**80.18 / 79.80**	46.45	**46.07**	**0.644**	**0.627**
Random Missing Rate $r = 0.1$							Random Missing Rate $r = 0.6$						
MISA	82.28 / 82.21	80.79 / 81.28	51.34	50.13	0.598	0.722	MISA	72.30 / 65.55	62.12 / 54.64	33.30	36.16	0.875	0.415
Self-MM	**83.79** / 83.03	**83.23** / 82.43	53.18	51.80	0.564	0.725	Self-MM	73.93 / 69.33	66.76 / 63.01	41.75	41.47	0.762	0.401
MMIM	81.09 / 82.00	80.15 / 81.57	51.19	49.09	0.602	0.696	MMIM	73.16 / 68.83	65.43 / 63.09	37.48	37.13	0.808	0.402
TFR-Net	83.31 / 82.92	82.40 / 82.25	45.82	52.29	0.573	0.715	TFR-Net	74.05 / 68.80	67.07 / 62.51	28.33	43.88	0.762	0.397
CENET	82.41 / 83.75	82.34 / 83.42	**54.23**	**52.83**	**0.556**	0.739	CENET	75.39 / 70.50	70.86 / 65.27	44.64	**44.45**	0.749	0.446
ALMT	82.84 / 84.14	83.04 / 84.23	52.38	50.98	0.583	0.718	ALMT	76.26 / 74.98	76.71 / 75.44	36.30	35.99	0.710	0.395
LNLN	82.73 / 83.32	82.91 / 83.66	51.25	49.96	0.591	0.712	LNLN	76.50 / 73.82	78.33 / 75.03	44.00	43.52	0.736	0.471
TF-Mamba	82.68 / 83.16	82.69 / 83.03	52.07	50.53	0.570	0.730	TF-Mamba	75.77 / 75.89	75.59 / 75.63	44.73	43.96	0.709	0.534
P-RMF	82.79 / 83.98	82.94 / **84.37**	51.45	49.04	**0.556**	**0.748**	P-RMF	77.44 / 76.11	78.97 / 77.58	43.12	42.26	0.703	0.545
EASE	83.06 / **84.26**	83.15 / 84.30	52.55	50.65	0.561	0.740	**EASE**	**78.65 / 76.53**	**79.62 / 77.69**	**45.36**	44.23	**0.675**	**0.570**
Random Missing Rate $r = 0.2$							Random Missing Rate $r = 0.7$						
MISA	79.93 / 77.84	76.88 / 75.56	47.66	47.24	0.659	0.674	MISA	71.71 / 64.28	60.65 / 51.82	31.21	34.54	0.906	0.344
Self-MM	**82.33** / 80.84	81.17 / 79.76	50.51	49.44	0.604	0.678	Self-MM	72.55 / 66.79	63.45 / 58.05	40.12	39.93	0.786	0.329
MMIM	79.66 / 79.93	77.68 / 79.08	47.99	46.27	0.642	0.653	MMIM	72.26 / 66.89	63.26 / 58.90	35.47	35.25	0.834	0.341
TFR-Net	81.61 / 80.47	79.99 / 79.29	40.13	**51.04**	0.604	0.672	TFR-Net	72.77 / 66.64	64.02 / 58.32	26.92	42.91	0.786	0.322
CENET	81.62 / 81.46	81.17 / 80.78	**51.85**	50.72	0.590	0.698	CENET	73.39 / 67.50	67.02 / 59.88	44.03	**43.93**	0.776	0.384
ALMT	81.65 / 82.71	81.83 / 82.82	47.82	46.61	0.607	0.669	ALMT	73.98 / 71.62	74.54 / 72.24	34.95	34.78	0.743	0.315
LNLN	81.68 / 81.70	81.89 / 81.95	49.95	48.75	0.616	0.677	LNLN	74.74 / 71.55	77.40 / 73.49	42.56	42.22	0.762	0.408
TF-Mamba	81.84 / 82.55	81.83 / 82.40	50.50	49.17	0.588	0.710	TF-Mamba	73.49 / 73.50	73.19 / 73.26	43.40	42.78	0.747	**0.469**
P-RMF	82.25 / 82.97	**82.58 / 83.37**	49.35	47.91	**0.576**	**0.722**	P-RMF	75.87 / 74.64	**78.12** / 75.88	42.41	41.73	0.733	0.481
EASE	82.11 / **83.09**	82.20 / 83.15	51.09	49.23	0.584	0.719	**EASE**	**76.73 / 75.97**	77.86 / **76.40**	**44.29**	42.71	**0.708**	**0.524**
Random Missing Rate $r = 0.3$							Random Missing Rate $r = 0.8$						
MISA	77.28 / 73.32	72.25 / 68.91	43.40	43.99	0.724	0.615	MISA	71.30 / 63.43	59.69 / 49.95	29.51	33.29	0.927	0.267
Self-MM	79.99 / 77.63	77.74 / 75.69	48.07	47.23	0.653	0.610	Self-MM	71.83 / 65.07	61.49 / 54.44	38.78	38.69	0.805	0.259
MMIM	77.79 / 77.08	74.49 / 75.46	44.73	43.25	0.690	0.597	MMIM	71.57 / 64.97	61.45 / 54.76	33.71	33.64	0.858	0.269
TFR-Net	79.29 / 77.48	76.52 / 75.43	38.34	**48.75**	0.650	0.604	TFR-Net	71.95 / 65.05	61.82 / 54.91	27.70	42.23	0.807	0.241
CENET	80.02 / 78.65	78.94 / 77.34	**49.37**	48.49	0.636	0.640	CENET	72.16 / 65.88	64.67 / 56.80	42.74	**42.71**	0.798	0.316
ALMT	79.94 / 80.94	80.20 / 81.15	44.05	43.04	0.632	0.598	ALMT	71.48 / 68.15	72.28 / 69.12	34.09	34.01	0.774	0.231
LNLN	80.45 / 80.11	80.91 / 80.44	48.40	47.36	0.648	0.629	LNLN	72.86 / 68.62	76.80 / 71.83	40.97	40.76	0.791	0.325
TF-Mamba	81.22 / 80.82	81.10 / 80.58	49.00	47.89	0.613	0.675	TF-Mamba	71.60 / 70.34	71.27 / 70.16	40.93	40.37	0.786	0.408
P-RMF	80.88 / 81.26	81.40 / 81.86	47.78	45.95	0.611	0.683	P-RMF	74.46 / 70.75	**77.91** / 73.03	41.00	40.46	0.764	0.401
EASE	**81.45 / 82.12**	**81.62 / 82.39**	49.78	47.74	**0.609**	**0.684**	**EASE**	**75.08 / 73.26**	76.14 / **74.37**	**43.14**	41.38	**0.731**	**0.476**
Random Missing Rate $r = 0.4$							Random Missing Rate $r = 0.9$						
MISA	75.04 / 70.46	67.93 / 64.02	39.53	40.87	0.780	0.561	MISA	71.07 / 62.95	59.12 / 48.80	28.03	32.29	0.941	0.180
Self-MM	78.09 / 75.02	74.48 / 72.01	45.04	44.40	0.694	0.554	Self-MM	71.24 / 63.85	59.72 / 51.32	37.50	37.46	0.821	0.188
MMIM	76.15 / 74.56	71.40 / 71.98	41.86	40.84	0.732	0.542	MMIM	71.10 / 63.69	59.99 / 51.26	32.67	32.61	0.877	0.197
TFR-Net	77.65 / 74.74	73.71 / 71.67	35.76	46.70	0.688	0.548	TFR-Net	71.34 / 63.64	59.99 / 52.02	25.12	41.73	0.820	0.175
CENET	78.57 / 76.03	76.75 / 73.87	47.74	**47.12**	0.678	0.587	CENET	70.42 / 64.14	62.33 / 54.27	**42.08**	**42.08**	0.814	0.254
ALMT	79.16 / 79.40	79.50 / 79.68	41.21	40.40	0.651	0.536	ALMT	68.65 / 61.41	69.83 / 63.32	34.40	34.40	0.810	0.138
LNLN	79.70 / 78.49	80.46 / 78.98	46.88	45.99	0.673	0.592	LNLN	71.51 / 64.83	77.52 / 70.60	40.19	40.10	0.820	0.221
TF-Mamba	80.02 / 80.02	79.80 / 79.80	**47.80**	46.73	0.639	0.638	TF-Mamba	68.68 / 64.75	68.18 / 64.87	37.56	37.24	0.851	0.291
P-RMF	79.76 / 79.97	80.58 / 80.74	46.73	45.59	0.631	**0.653**	P-RMF	72.59 / 67.86	**77.95** / 71.51	39.90	39.62	0.805	0.289
EASE	**80.62 / 80.68**	**81.09 / 80.93**	47.51	46.98	**0.622**	0.650	**EASE**	**73.24 / 70.79**	74.57 / **72.03**	41.90	40.19	**0.778**	**0.342**

Table 8: Details of robust comparison on MOSEI with different random missing rates. Refer to Sec. 4.3.



(a) Distribution Alignment

(b) Sentiment Predictive Consistency

Figure 9: Visualization of the MIPC module's impact on MOSI. Best viewed in color. Refer to Sec. 4.5.

Method	Acc-2↑	F1↑	Acc-3↑	Acc-5↑	MAE↓	Corr↑	Method	Acc-2↑	F1↑	Acc-3↑	Acc-5↑	MAE↓	Corr↑
Random Missing Rate $r = 0$							Random Missing Rate $r = 0.5$						
MISA	78.19	77.22	63.38	40.55	0.449	0.576	MISA	71.26	64.16	54.78	30.56	0.552	0.367
Self-MM	78.26	78.00	**64.92**	**40.77**	**0.421**	**0.584**	Self-MM	71.41	67.11	53.90	32.02	0.517	0.390
MMIM	75.42	73.10	60.69	37.42	0.475	0.528	MMIM	68.49	64.81	52.37	33.41	0.553	0.336
TFR-Net	69.15	58.44	54.12	33.85	0.562	0.254	TFR-Net	67.47	58.66	52.37	24.65	0.685	0.171
CENET	68.71	57.82	54.05	23.85	0.578	0.137	CENET	68.71	57.92	54.05	23.12	0.588	0.107
ALMT	75.64	76.27	54.78	23.41	0.527	0.536	ALMT	68.27	71.22	47.12	18.38	0.563	0.395
LNLN	75.93	**79.89**	63.97	38.66	0.458	0.570	LNLN	72.72	**78.77**	57.70	36.40	0.513	0.412
TF-Mamba	**79.65**	78.92	61.93	37.86	0.441	0.548	TF-Mamba	**75.71**	73.61	**58.64**	**37.20**	0.495	0.424
P-RMF	78.34	79.69	60.61	38.95	0.441	0.550	P-RMF	73.61	74.83	54.83	34.79	0.495	0.404
EASE	78.15	78.32	62.83	38.45	0.424	0.560	**EASE**	74.58	74.77	58.26	36.11	**0.482**	**0.454**
Random Missing Rate $r = 0.1$							Random Missing Rate $r = 0.6$						
MISA	77.39	75.82	63.02	38.88	0.461	0.561	MISA	70.46	61.81	53.97	27.72	0.578	0.286
Self-MM	77.32	76.76	**63.53**	**40.26**	**0.433**	**0.563**	Self-MM	70.02	64.21	51.86	29.10	0.548	0.313
MMIM	74.25	72.08	60.90	37.27	0.473	0.529	MMIM	67.91	63.86	49.31	29.18	0.578	0.270
TFR-Net	68.85	59.38	53.25	30.12	0.596	0.203	TFR-Net	67.03	58.30	52.59	24.80	0.696	0.157
CENET	68.57	57.36	53.98	22.83	0.580	0.136	CENET	69.00	58.64	53.69	22.46	0.592	0.102
ALMT	74.40	75.19	55.14	22.10	0.530	0.537	ALMT	66.81	70.69	43.69	18.67	0.574	0.322
LNLN	76.29	**80.07**	62.73	38.51	0.458	0.562	LNLN	71.55	**79.56**	54.63	33.70	0.535	0.352
TF-Mamba	**79.65**	78.79	62.36	36.98	0.445	0.550	TF-Mamba	73.09	69.44	54.92	32.82	0.523	0.370
P-RMF	77.24	78.88	59.52	37.72	0.454	0.530	P-RMF	72.12	73.32	53.05	33.92	0.515	0.383
EASE	77.61	77.52	61.81	37.86	0.438	0.538	**EASE**	**73.90**	74.29	**57.10**	**35.37**	**0.501**	**0.428**
Random Missing Rate $r = 0.2$							Random Missing Rate $r = 0.7$						
MISA	74.33	71.70	59.23	38.15	0.489	0.490	MISA	69.95	59.54	52.52	24.87	0.601	0.167
Self-MM	74.98	73.71	61.71	38.37	0.464	0.500	Self-MM	69.58	62.28	50.62	25.53	0.571	0.198
MMIM	72.36	69.80	57.33	37.27	0.504	0.460	MMIM	66.89	62.23	46.53	28.59	0.595	0.190
TFR-Net	68.64	59.74	53.61	29.03	0.619	0.191	TFR-Net	67.18	58.15	52.30	23.78	0.707	0.163
CENET	68.57	57.64	54.20	22.25	0.583	0.132	CENET	67.69	57.87	53.32	21.81	0.599	0.070
ALMT	72.65	73.90	53.17	21.08	0.541	0.485	ALMT	65.57	70.27	38.66	18.02	0.586	0.218
LNLN	74.76	**78.53**	**61.78**	**38.88**	0.474	0.513	LNLN	69.73	**79.37**	51.64	30.27	0.558	0.261
TF-Mamba	**78.77**	77.88	61.05	38.29	0.459	0.507	TF-Mamba	71.77	67.46	48.80	28.45	0.570	0.248
P-RMF	76.23	77.46	59.30	37.05	0.460	0.512	P-RMF	71.58	72.22	52.52	32.23	0.531	0.369
EASE	76.99	77.19	60.79	37.32	**0.447**	**0.521**	**EASE**	**72.76**	73.65	**55.04**	**33.94**	**0.520**	**0.395**
Random Missing Rate $r = 0.3$							Random Missing Rate $r = 0.8$						
MISA	74.11	70.40	59.30	36.40	0.505	0.464	MISA	69.37	57.82	52.22	22.69	0.610	0.092
Self-MM	74.76	72.85	59.81	37.93	0.474	0.487	Self-MM	69.51	60.68	50.77	22.03	0.585	0.138
MMIM	72.36	69.52	58.06	37.71	0.512	0.436	MMIM	65.28	60.53	44.35	22.32	0.607	0.145
TFR-Net	68.42	59.88	52.30	27.64	0.640	0.182	TFR-Net	67.54	57.55	52.74	22.97	0.721	0.100
CENET	68.42	57.41	54.05	21.44	0.578	0.175	CENET	67.47	58.44	52.15	21.73	0.599	0.074
ALMT	72.06	73.64	50.62	20.35	0.546	0.469	ALMT	64.19	69.64	34.06	18.60	0.597	0.133
LNLN	74.25	**78.60**	**60.98**	38.37	0.478	0.509	LNLN	69.58	**80.23**	50.47	27.94	0.580	0.183
TF-Mamba	75.93	74.72	58.42	**39.82**	0.468	0.485	TF-Mamba	**71.55**	66.58	46.39	26.48	0.607	0.139
P-RMF	75.66	75.89	56.89	36.89	0.475	0.483	P-RMF	69.51	70.35	49.89	31.07	0.568	0.269
EASE	**76.28**	76.78	60.05	36.84	**0.454**	**0.506**	**EASE**	71.44	72.28	**53.43**	**32.69**	**0.542**	**0.341**
Random Missing Rate $r = 0.4$							Random Missing Rate $r = 0.9$						
MISA	72.87	67.52	57.33	34.86	0.523	0.436	MISA	69.22	57.01	52.95	20.64	0.617	0.041
Self-MM	73.30	70.36	58.28	34.57	0.482	0.479	Self-MM	68.92	58.32	52.15	22.17	0.586	0.111
MMIM	69.95	66.49	55.36	34.57	0.533	0.399	MMIM	65.72	59.64	42.67	20.35	0.610	0.096
TFR-Net	67.91	59.16	51.86	25.31	0.664	0.176	TFR-Net	69.08	57.71	**53.76**	23.05	0.721	0.088
CENET	68.49	57.68	54.12	22.54	0.583	0.141	CENET	65.72	58.18	48.07	20.86	0.609	-0.002
ALMT	70.75	72.97	49.45	19.91	0.549	0.470	ALMT	66.23	73.76	26.91	19.47	0.596	0.076
LNLN	73.81	**78.82**	60.03	37.49	0.491	**0.481**	LNLN	68.64	**80.42**	47.48	26.19	0.591	0.127
TF-Mamba	75.49	73.72	**60.39**	**40.04**	**0.470**	0.477	TF-Mamba	65.21	60.87	42.23	26.70	0.648	0.114
P-RMF	75.32	76.30	55.80	36.54	0.482	0.472	P-RMF	66.77	67.54	45.08	29.10	0.581	0.168
EASE	**75.61**	76.04	59.47	36.70	**0.470**	**0.493**	**EASE**	**69.91**	70.50	51.97	**31.02**	**0.573**	**0.279**

Table 9: Detailed robustness comparison on SIMS under different intra-modality missing rates. Refer to Sec. 4.3.

Fix u	Method	MOSI						MOSEI					
		Acc-2	F1	Acc-5	Acc-7	MAE	Corr	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
$\{l\}$	MISA	81.24 / 83.28	81.17 / 83.29	48.40	43.25	0.768	0.776	83.97 / 84.88	83.64 / 84.64	53.80	51.67	0.558	0.756
	Self-MM	82.80 / 85.06	82.82 / 85.04	52.77	42.71	0.722	0.789	84.74 / 85.32	84.67 / 85.06	55.43	53.67	0.535	0.761
	MMIM	81.29 / 83.48	81.17 / 83.42	51.07	44.56	0.748	0.777	81.14 / 83.79	80.96 / 83.61	53.07	50.70	0.579	0.722
	CENET	81.54 / 83.08	81.54 / 83.06	50.39	43.05	0.750	0.785	80.66 / 84.64	81.23 / 84.71	54.22	52.71	0.545	0.760
	TETFN	81.05 / 82.57	81.04 / 82.62	51.36	44.12	0.719	0.794	83.79 / 84.83	83.85 / 84.80	55.22	53.46	0.549	0.747
	TFR-Net	79.98 / 83.08	79.82 / 83.01	43.44	36.00	0.838	0.758	83.49 / 83.97	83.18 / 83.91	54.54	53.13	0.578	0.726
	ALMT	79.83 / 84.55	80.65 / 84.76	41.93	36.49	0.864	0.767	69.11 / 67.14	72.15 / 71.77	33.57	30.64	0.560	0.748
	LNLN	82.26 / 84.86	82.48 / 85.12	52.04	45.10	0.760	0.772	82.69 / 84.10	82.86 / 84.39	51.38	50.10	0.609	0.730
	P-RMF	80.90 / 81.01	80.89 / 81.36	47.46	42.57	0.777	0.759	81.52 / 82.26	82.11 / 81.91	50.46	49.09	0.562	0.757
	EASE	81.24 / 83.09	80.28 / 83.10	48.33	43.74	0.761	0.752	81.69 / 83.46	81.72 / 83.61	51.76	49.22	0.560	0.738
$\{a\}$	MISA	50.87 / 48.98	43.51 / 41.79	15.40	15.45	1.427	0.169	71.02 / 62.85	58.99 / 48.51	26.89	31.57	0.936	0.105
	Self-MM	51.75 / 57.77	35.44 / 42.31	19.24	16.47	1.386	0.072	71.02 / 62.85	58.99 / 48.51	36.55	36.55	0.838	0.101
	MMIM	48.64 / 53.05	36.52 / 43.80	17.40	17.93	1.366	0.155	70.94 / 62.88	59.00 / 48.91	32.12	32.12	0.891	0.143
	CENET	51.80 / 57.67	35.71 / 42.26	17.54	17.54	1.387	0.118	67.49 / 61.78	59.63 / 51.87	36.59	36.59	0.838	0.112
	TETFN	55.25 / 57.77	39.32 / 42.31	21.14	21.19	1.403	0.097	71.02 / 62.81	58.99 / 48.51	41.36	41.36	0.840	0.017
	TFR-Net	55.25 / 57.77	39.32 / 42.31	25.07	22.50	1.486	0.154	71.02 / 62.47	58.99 / 51.16	41.36	41.36	0.839	0.039
	ALMT	55.10 / 56.45	66.11 / 67.09	19.77	19.29	1.394	0.136	57.01 / 54.28	70.35 / 69.52	21.54	21.54	0.874	-0.086
	LNLN	49.03 / 52.18	56.84 / 58.89	17.68	18.80	1.427	0.075	71.02 / 62.85	83.06 / 77.19	38.41	38.41	0.853	0.052
	P-RMF	55.03 / 56.85	71.72 / 71.44	20.99	20.55	1.367	0.107	71.02 / 62.85	83.06 / 75.91	41.74	41.25	0.838	0.115
	EASE	59.44 / 61.15	66.72 / 68.89	25.82	23.34	1.317	0.192	70.91 / 67.38	82.79 / 77.56	41.58	40.02	0.813	0.179
$\{v\}$	MISA	55.20 / 55.29	49.24 / 49.53	15.60	15.55	1.419	0.098	71.02 / 62.85	58.99 / 48.51	26.76	31.17	0.961	0.121
	Self-MM	54.03 / 57.77	43.31 / 42.31	19.29	16.86	1.381	0.130	71.02 / 63.01	58.99 / 49.18	36.61	36.61	0.831	0.144
	MMIM	48.20 / 52.24	31.53 / 39.57	17.40	17.59	1.389	0.024	70.94 / 62.31	59.28 / 54.32	31.92	31.91	0.886	0.182
	CENET	51.85 / 57.67	35.65 / 42.26	17.40	17.40	1.387	0.110	64.21 / 60.69	56.68 / 51.22	41.40	41.40	0.823	0.236
	TETFN	55.25 / 57.77	39.32 / 42.31	21.14	21.19	1.403	0.098	71.00 / 62.82	58.98 / 48.51	41.31	41.31	0.832	0.144
	TFR-Net	55.59 / 58.03	40.54 / 43.37	23.96	21.38	1.554	0.107	71.02 / 62.81	58.99 / 48.75	41.38	41.37	0.829	0.161
	ALMT	54.96 / 56.35	65.86 / 66.89	19.24	19.24	1.397	0.104	57.28 / 54.28	70.09 / 69.52	21.57	21.57	0.870	0.091
	LNLN	49.03 / 52.18	56.84 / 58.89	17.68	18.80	1.427	0.072	71.02 / 62.85	83.06 / 77.19	34.72	34.72	0.900	0.145
	P-RMF	54.96 / 56.01	70.72 / 70.32	20.26	19.83	1.368	0.099	70.52 / 61.77	82.14 / 73.19	33.68	33.59	0.828	0.209
	EASE	58.92 / 60.34	66.51 / 68.59	25.11	22.87	1.325	0.156	70.79 / 66.91	82.45 / 76.80	41.20	39.75	0.809	0.274
$\{l, a\}$	MISA	81.00 / 82.37	80.99 / 82.43	47.09	43.29	0.777	0.777	83.99 / 85.25	83.76 / 85.11	53.80	51.76	0.550	0.757
	Self-MM	82.80 / 85.11	82.82 / 85.09	55.47	42.76	0.722	0.789	84.62 / 85.20	84.59 / 84.95	55.47	53.68	0.533	0.761
	MMIM	81.83 / 83.54	81.78 / 83.52	49.85	44.75	0.740	0.778	81.31 / 83.71	81.09 / 83.53	52.86	50.45	0.579	0.722
	CENET	81.49 / 83.08	81.48 / 83.06	50.39	43.20	0.748	0.785	81.04 / 84.99	81.56 / 85.03	54.48	52.95	0.544	0.761
	TETFN	81.10 / 82.62	81.09 / 82.67	51.46	44.22	0.719	0.794	83.78 / 84.84	83.86 / 84.81	55.34	53.55	0.544	0.761
	TFR-Net	81.10 / 83.54	80.95 / 83.41	45.87	37.85	0.799	0.760	83.48 / 83.97	83.18 / 83.93	54.45	52.94	0.577	0.726
	ALMT	79.98 / 84.81	80.76 / 84.81	41.93	36.39	0.863	0.768	69.41 / 66.81	72.23 / 71.72	32.95	29.80	0.559	0.747
	LNLN	82.26 / 84.91	82.48 / 85.17	51.99	45.05	0.759	0.772	83.41 / 84.09	83.76 / 84.43	51.44	50.16	0.577	0.729
	P-RMF	81.05 / 82.16	81.03 / 82.10	48.69	43.29	0.776	0.760	83.58 / 84.62	83.60 / 84.61	50.42	49.04	0.565	0.757
	EASE	81.69 / 83.70	81.95 / 83.87	49.02	44.23	0.752	0.775	83.44 / 84.79	83.71 / 85.06	52.66	50.15	0.548	0.739
$\{l, v\}$	MISA	81.58 / 83.74	81.49 / 83.72	48.44	43.49	0.762	0.776	83.91 / 84.94	83.46 / 84.62	53.91	51.59	0.561	0.759
	Self-MM	83.24 / 85.22	83.26 / 85.19	52.33	42.81	0.720	0.790	84.70 / 85.31	84.64 / 85.06	55.66	53.91	0.532	0.763
	MMIM	81.15 / 83.08	81.02 / 83.00	49.52	45.39	0.744	0.777	81.51 / 83.72	81.29 / 83.60	53.26	51.01	0.572	0.726
	CENET	81.54 / 83.08	81.54 / 83.06	50.39	43.05	0.750	0.785	82.08 / 85.37	82.44 / 85.35	55.98	54.34	0.531	0.770
	TETFN	81.00 / 82.52	80.99 / 82.57	51.31	44.07	0.719	0.794	83.99 / 85.08	84.05 / 85.05	55.94	54.15	0.543	0.753
	TFR-Net	80.52 / 82.72	80.45 / 82.71	45.48	37.71	0.821	0.759	84.65 / 84.95	84.32 / 84.69	55.34	53.75	0.550	0.745
	ALMT	79.88 / 84.60	80.70 / 84.81	41.84	36.49	0.865	0.767	70.05 / 67.13	72.59 / 71.76	33.55	30.51	0.535	0.556
	LNLN	82.21 / 84.86	82.43 / 85.12	52.04	45.05	0.760	0.772	83.09 / 84.39	83.34 / 84.71	51.92	50.63	0.610	0.736
	P-RMF	80.90 / 82.01	80.89 / 81.94	48.40	43.15	0.777	0.759	83.52 / 85.20	83.49 / 85.17	51.87	49.41	0.559	0.765
	EASE	81.50 / 83.44	81.76 / 83.60	49.37	44.19	0.740	0.778	83.32 / 84.65	83.40 / 84.98	52.70	50.27	0.549	0.742
$\{a, v\}$	MISA	53.74 / 52.18	49.42 / 48.09	15.65	15.60	1.422	0.171	71.02 / 62.85	58.99 / 48.51	26.75	30.93	0.946	0.128
	Self-MM	54.37 / 57.77	43.81 / 42.31	19.29	16.86	1.381	0.129	71.02 / 63.03	58.99 / 49.22	36.59	36.59	0.832	0.152
	MMIM	49.61 / 54.67	37.63 / 46.95	17.20	17.44	1.363	0.153	71.01 / 62.71	59.43 / 54.70	31.97	31.96	0.886	0.195
	CENET	51.80 / 57.67	35.71 / 42.26	17.54	17.54	1.387	0.118	69.38 / 63.23	60.67 / 52.41	41.70	41.70	0.821	0.238
	TETFN	55.25 / 57.77	39.32 / 42.31	21.14	21.19	1.403	0.097	71.01 / 62.82	58.98 / 48.53	41.31	41.31	0.831	0.148
	TFR-Net	55.30 / 57.82	39.43 / 42.43	25.95	23.13	1.563	0.167	71.02 / 62.85	58.99 / 48.88	41.43	41.40	0.829	0.163
	ALMT	55.05 / 56.40	66.01 / 66.99	19.68	19.29	1.394	0.139	57.38 / 54.28	70.01 / 69.52	21.58	21.58	0.867	0.071
	LNLN	49.03 / 52.18	56.84 / 58.89	17.68	18.80	1.427	0.075	71.02 / 62.85	83.06 / 77.19	39.10	39.10	0.847	0.156
	P-RMF	55.95 / 58.42	71.98 / 73.11	20.85	20.41	1.366	0.109	70.68 / 63.34	82.16 / 76.88	37.63	33.59	0.822	0.211
	EASE	59.89 / 60.85	66.74 / 68.61	25.37	23.19	1.314	0.184	71.02 / 67.57	82.82 / 77.73	41.82	39.23	0.810	0.252

Table 10: Generalization comparison on MOSI and MOSEI with inter-modality missing. Note: The smaller MAE indicates the better performance.