Proxy-Driven Robust Multimodal Sentiment Analysis with Incomplete Data

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, Ning An
School of Computer Science and Information Engineering, Hefei University of Technology
zhuaoqiang@mail.hfut.edu.cn, {jsjxhumin, xh_wang, jiaoyun, ymtang}@hfut.edu.cn, ning.g.an@acm.org

Abstract

Multimodal Sentiment Analysis (MSA) with incomplete data has gained significant attention recently. Existing studies focus on optimizing model structures to handle modality missingness, but models still face challenges in robustness when dealing with uncertain missingness. To this end, we propose a data-centric robust multimodal sentiment analysis method, Proxy-Driven Robust Multimodal Fusion (P-RMF). First, we map unimodal data to the latent space of Gaussian distributions to capture core features and structure, thereby learn stable modality representation. Then, we combine the quantified modality intrinsic uncertainty to learn stable multimodal joint representation (i.e., proxy modality), which is further enhanced through multi-layer dynamic cross-modal injection to increase its diversity. Extensive experimental results show that P-RMF outperforms existing models in noise resistance and achieves state-of-the-art performance on multiple benchmark datasets. Code will be available at https://github.com/aoqzhu/P-RMF.

1 Introduction

Multimodal Sentiment Analysis (MSA) integrates information from multiple modalities to understand and recognize human emotions (Singh et al., 2024). By fusing complementary information, multimodal learning generates richer joint representation (Xu et al., 2023). However, in real-world scenarios, modality data is often missing due to factors such as background noise, sensor limitations, and privacy concerns. Data incompleteness significantly reduces the effectiveness of models trained on complete data (Wang et al., 2023b).

In recent years, many studies (Wang et al., 2023c; Zhang et al., 2024; Sun et al., 2023b; Li et al., 2024b) have proposed influential solutions to the missing data problem in MSA. These methods can be broadly categorized into two main types: reconstruction-based methods (Lian et al., 2023; Sun et al., 2023b; Zhao et al., 2021) and joint representation-based methods (Li et al., 2024a,b; Kim and Kim, 2024). Although these methods have made significant progress in addressing data incompleteness, models focusing on model structure optimization still face challenges in performance and robustness when handling data with uncertain missingness. These challenges mainly stem from distributional discrepancies in incomplete data and the epistemic uncertainty of deep learning models (Kendall and Gal, 2017). Moreover, existing models for incomplete data often fail to balance the effective processing of complete data.

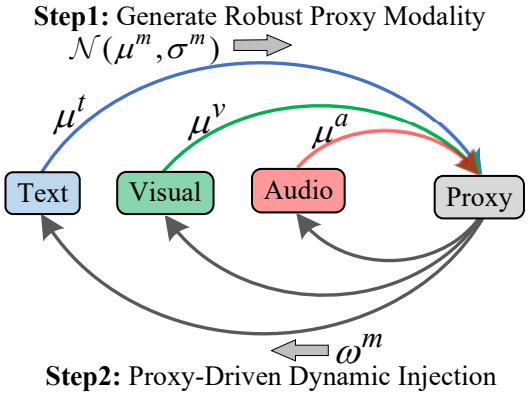

Figure 1: The core steps of P-RMF.

The randomly missing incomplete data further increases the arbitrary uncertainty of the data and the epistemic uncertainty of the model (Kendall and Gal, 2017). Therefore, we explore whether a data-centric method, focusing on the core features and structures of the data, can enhance model robustness to input variations and noise. Inspired by probability distributions (Ho et al., 2020; Chen et al., 2022; Yang et al., 2024), we argue that semantically related data, even across different modalities, should exhibit similar distributions in the latent Gaussian space despite varying noise levels. This latent space constructs regularized, compact, and semantically consistent representations, abstracting

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22123–22138

high-level features while weakening task-irrelevant details, thereby enhancing robustness to input variations and noise. Given the stable and semantically consistent unimodal representation, should we disregard the intrinsic uncertainty of the data? Prior studies (Sun et al., 2023a; Yu et al., 2020) highlight the unique characteristics of unimodal data and the varying contributions of different modalities to sentiment recognition. We argue that the uncertainty arising from missing noise in heterogeneous data reflects inter-modal information discrepancies, embodying the distinct properties of each modality. Thus, we quantify this uncertainty by measuring feature space distribution discrepancies and integrate it into multimodal fusion to learn dynamic and robust representation.

Based on the above analysis, we propose Proxy-Driven Robust Multimodal Fusion (P-RMF) for robust sentiment analysis under uncertain missing data. As shown in Fig. 1, P-RMF consists of two core steps: generate robust proxy modality and proxy-driven dynamic injection. Specifically, we employ variational inference and design an independent Variational Autoencoder (VAE) for each modality. The encoder maps data into a Gaussian distributions latent space, learning its mean and variance, while the reparameterization trick enables sampling of latent variables as compact representation of missing data. The decoder then reconstructs the complete input, learning to recover missing information. The variance captures modality uncertainty, whereas the mean provides a stable representation. By integrating both, We learn a consistent and stable joint modality representation (i.e., proxy modality). Finally, with proxy modality as the dominant, combined with quantized uncertainty, we enhance its diversity by multi-layer dynamic cross-modal injection. The main contributions of this paper are as follows:

- We propose a data-centric robust multimodal sentiment analysis method to address the robustness problem under uncertain missing data. This method learns stable unimodal representations from missing data and incorporates quantization uncertainty to guide the learning of dynamic and robust multimodal representation.
- We propose a robust proxy modality generation module that learns stable data representations in a latent space based on Gaussian distributions and generates a robust joint proxy modality representation by incorporating quantized uncertainty.
- We propose a proxy modality-driven dynamic injection module that incorporates quantized modality uncertainty, iteratively injecting modality-specific semantics with varying weights into the proxy modality to enhance the diversity of the joint representation.

2 Related Work

Multimodal Sentiment Analysis (MSA) can be categorized into context-based MSA and noise-aware robust MSA based on modeling methods. Early context-based MSA studies assumed that all data are complete and available during both training and inference phases (Han et al., 2021; Qian et al., 2023; Yu et al., 2023; Li et al., 2023b; Zhu et al., 2024; Zhang et al., 2023). These studies primarily focused on learning unified multimodal representations by analyzing intra- and inter-modal contextual relationships. Despite the progress made by these methods, incomplete data can significantly reduce the effectiveness of models trained on complete data.

In recent years, many studies (Li et al., 2023a, 2024a; Zeng et al., 2022; Zhang et al., 2024; Sun et al., 2023b; Li et al., 2024b) have attempted to solve the missing data problem in MSA by methods such as data reconstruction networks and joint representation learning. For example, EMT-DLFR (Sun et al., 2023b) encouraged the model to learn semantic information from missing data by performing low-level feature reconstruction. These reconstruction-based methods focus on complementing missing modalities using existing modalities, but the quality of the reconstructed data is often difficult to guarantee and is less interpretable. Therefore, UMDF (Li et al., 2024a) and CorrKD (Li et al., 2024b) introduced a knowledge distillation mechanism to learn joint multimodal representations by optimizing the fusion strategy and model architecture. Further, LNLN (Zhang et al., 2024) shows that stabilizing the text-dominant modality under varying levels of missing noise improves model robustness, but ensuring the integrity and quality of the text modality is challenging in multimodal data with arbitrary missing uncertainties.

Inspired by LNLN (Zhang et al., 2024), we argue that the stable dominant modality need not be confined to the original modality (e.g., text or

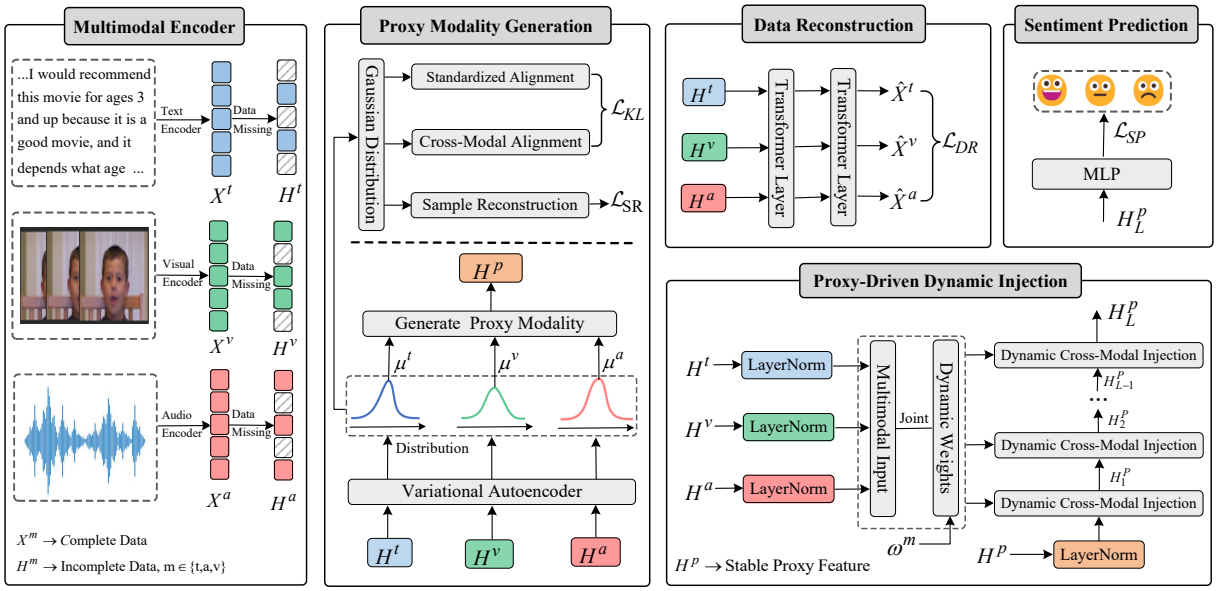

Figure 2: The framework of the P-RMF.

visual), but can instead be represented by the semantic consistency of different modalities in the high-level feature space. Therefore, we propose a Proxy-Driven Robust Multimodal Fusion (P-RMF).

3 Methodology

3.1 Overall Framework

Fig. 2 shows the key components and workflow of the proposed Proxy-Driven Robust Multimodal Fusion (P-RMF) method. P-RMF consists of five main components: Multimodal Encoder, Proxy Modality Generation (PMG), Proxy-Driven Dynamic Injection (PDDI), Data Reconstruction (DR), and Sentiment Prediction. During training, P-RMF jointly inputs complete data and randomly missing incomplete data. With the joint constraint of KL divergence loss and variational sample reconstruction loss, PMG generates a consistent and robust proxy modality representation by combining the quantized intrinsic modality uncertainty. The PDDI incorporates quantized uncertainty to enhance the diversity of proxy modality representations through dynamic cross-modal injection iterations in the multi-layer. In addition, to complement the missing fine-grained sentiment semantics, P-RMF designs a reconstructor for each modality to reconstruct the missing data. During testing, the trained P-RMF is applied to scenarios involving random intra-modal and inter-modal missing data to evaluate its robustness and effectiveness.

3.2 Input Construction and Multimodal Encoder

We conduct experimental analysis on the MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018) and SIMS (Yu et al., 2020) datasets. To simulate random data missing scenarios, we follow the settings of previous work (Zhang et al., 2024), randomly erasing 0% to 100% of information in each modality for every sample. Specifically, 0 padding is applied to the erased portions of the visual and audio modalities, while the text modality is filled with the unknown word token [UNK].

For feature extraction, this study follows prior work (Zhang et al., 2024): text modality is encoded using BERT (Devlin et al., 2019), visual features are extracted via OpenFace (Baltrusaitis et al., 2018), and audio features are obtained using Librosa (Brian McFee et al., 2015). For a given multimodal input, $X^m \in \mathbb{R}^{L_m \times d_m}$ represents the complete multimodal sequence, while $H^m \in \mathbb{R}^{L_m \times d_m}$ denotes the incomplete sequence with random missing data. Here, $m \in \{t, a, v\}$ represents the modality type (i.e., text, audio, and visual), L_m denotes the sequence length, and d_m denotes the feature vector dimension.

3.3 Proxy Modality Generation

Previous studies have shown that maintaining the integrity of the dominant Modality significantly enhances the robustness of the model (Zhang et al., 2024). Inspired by probability distributions (Ho et al., 2020; Chen et al., 2022; Yang et al., 2024),

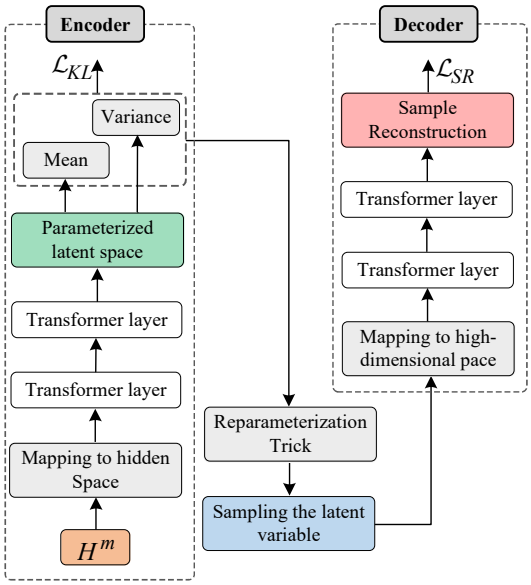

Figure 3: The overall architecture of the Variational Autoencoder.

we learn stable joint multimodal representation from Gaussian distributions of the data as the dominant Modality.

The key to Gaussian distribution analysis lies in how to map the data to the parameterized form of the mean and variance according to the task requirements. To achieve this, we have designed a Variational Autoencoder (VAE) for each modality, as shown in Fig. 3. In the VAE framework, the encoder learns the parameterized mean $u(h)$ and log-variance $\sigma^2(h)$ of the input data in the Gaussian-distributed latent space, and $\mathcal{N}(u(h), \sigma^2(h)I)$ represents the latent distribution of the data. The decoder then reconstructs the input data based on the generated latent variables z output from the encoder, recovering the core features.

Formally, the variational posterior distribution of the unimodal feature h can be expressed as $q(z|h) = \mathcal{N}\left[z\middle|u(h), \sigma^2(h)I\right]$. Thus, for the multimodal representation H_n^m of the n-th sample, the variational posterior distributions for different modalities can be expressed as:

$$q(z_n^m|h_n^m) \sim \mathcal{N}[z_n^m\middle|u(h_n^m), \sigma^2(h_n^m)I] \quad (1)$$

where $z_n^m, m \in \{t, v, a\}$ represents the latent variable for modality m of sample n, and I denotes the identity matrix. To make the learning of the latent space smoother and more stable, we introduce the Kullback-Leibler (KL) divergence as a regularization term to align the discrepancy between the posterior distribution $q(z_n^m|h_n^m)$ and the standard prior distribution $p(z_n^m) = \mathcal{N}(z_n^m|0, I)$. Additionally, we introduce cross-modal KL to align the distributional representations across modalities. The total KL constraint is computed as follows:

$$\begin{aligned}\mathcal{L}_{KL} = &\sum D_{KL}(q(z_n^{m_1}|h_n^{m_1})||q(z_n^{m_2}|h_n^{m_2})) \\ &+ \sum D_{KL}(q(z_n^m|h_n^m)||p(z_n^m))\end{aligned} \quad (2)$$

where $m \in M = \{t, v, a\}$ and $m_1 \neq m_2$. To effectively model the latent space, we employ the reparameterization trick, enabling sampling from the Gaussian distribution in the latent space. Specifically, the latent variable z is expressed as:

$$z = u(h) + \varepsilon \cdot \sigma(h),\ \varepsilon \sim \mathcal{N}(0, I) \quad (3)$$

where ε represents the noise sampled from the standard normal distribution. With this reparameterization method, we can convert the sampling process into a conductive operation, which allows the gradient to be efficiently transmitted via backpropagation to support model training. The latent variable z, as an abstract representation of the input data, captures its essential features, reflecting the effectiveness and validity of the latent space parameterization. By evaluating the reconstruction error between the reconstructed samples $\hat{h}_n^m = \text{Decoder}(z_n^m)$ and the complete data x, we ensure that the samples generated from the latent space closely approximate the complete data.

$$\mathcal{L}_{SR} = \frac{1}{N} \sum_{m \in M} \sum_{n=1}^{N} \left|\left| x_n^m - \hat{h}_n^m \right|\right|^2 \quad (4)$$

where x_n^m represents the complete data, $\hat{h}_n^m$ is the reconstruction of the incomplete input data, and N is the number of samples. Thus, the total loss for variational inference in multimodal data can be formulated as:

$$\mathcal{L}_{VAE} = \mathcal{L}_{KL} + \mathcal{L}_{SR} \quad (5)$$

The $\mathcal{L}_{VAE}$ constraint ensures effective and consistent stable representations in the latent space. The mean $u(h)$ of the Gaussian distribution represents stable features, while the variance $\sigma^2(h)$ reflects distribution uncertainty. Given the stable unimodal representation $u(h_n^m)$, we compute fusion weights ω_n^m for cross-modal representations based on modality uncertainty $\sigma(h_n^m)$, and apply them to the joint multimodal representation.

$$\begin{aligned}H^p &= \sum_{m \in M} \omega_n^m \cdot u(h_n^m), \\ \omega_n^m &= \frac{\exp(1/\sigma(h_n^m))}{\sum_{m \in M} \exp(1/\sigma(h_n^m))}\end{aligned} \quad (6)$$

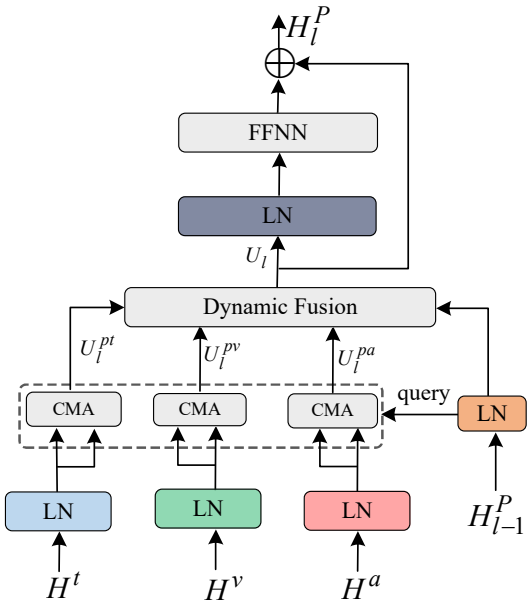

Figure 4: The framework of Proxy-Driven Dynamic Injection at layer l.

where H^p represents the stable joint multimodal representation, termed the proxy modality.

3.4 Proxy-Driven Dynamic Injection

The proxy modality H^p maintains the stability and semantic consistency of multimodal data by preserving the core features and structure of the data. However, the specific features of unimodal data reflect the diversity of multimodal representations. Therefore, we propose the PDDI. As shown in Fig. 2, the PDDI incorporates quantized uncertainty weights ω_n^m to enhance the diversity of proxy modality representation through dynamic cross-modal injection iterations in the multi-layer.

As in Fig. 4, we keep the multimodal features $H^m, m \in \{t, a, v\}$ unchanged and use Cross-Modal Attention (CMA) to inject multimodal information. Given the previous layer's output H_{l-1}^p and H^m, CMA is defined as:

$$U_l^{pm}=\text{softmax}(\frac{Q^p \cdot K^m}{\sqrt{d_t}}) \cdot V^m \tag{7}$$

where $Q^p = LN(H_{l-1}^p) \cdot W_Q^p$, $K^m = LN(H^m) \cdot W_K^m$, $V^a = LN(H^m) \cdot W_V^m$, LN is layer normalization. Then, $U_l^{pm}, m \in \{t, a, v\}$ is weighted by weights ω^m of each modality and combined with the previous layer's output H_{l-1}^p to obtain U_l.

$$U_l = LN(H_{l-1}^p) + \sum_{m \in M} \omega^m \cdot U_l^{pm} \tag{8}$$

Finally, U_l undergoes layer normalization, a feedforward neural network (FFNN), and a residual to obtain the output H_l^p of the current layer.

$$H_l^p = \text{FFNN}(LN(U_l)) + U_l \tag{9}$$

3.5 Data Reconstruction

To maximize the potential of available data and reduce information loss due to missing data, we design a data reconstructor R^m consisting of two Transformer layers for each modality to rebuild the missing information. For incomplete data H^m, we combine the complete data X^m with a mean square error loss function to optimize the performance of the reconstructor.

$$\mathcal{L}_{DR} = \frac{1}{N} \sum_{m \in M} \sum_{n=1}^{N} ||x_n^m - \text{R}^m(h_n^m)||^2 \tag{10}$$

The data reconstruction loss $\mathcal{L}_{DR}$ enhances the model's ability to capture sentiment information in the data by minimizing the discrepancies between the original and reconstructed features, thereby enhancing robustness.

3.6 Overall Learning Objectives

We feed the learned robust multimodal joint representation H_L^p into a Multi-Layer Perceptron (MLP) consisting of two fully connected layers and a ReLU activation function to output the sentiment prediction. The model is optimized using the L1 loss to minimize sentiment prediction error.

$$\mathcal{L}_{SP} = \frac{1}{N} \sum_{n=1}^{N} |\hat{y}_n - y_n| \tag{11}$$

where $\mathcal{L}_{SP}$ represents the sentiment prediction loss, $\hat{y}_n = \text{MLP}(H_L^p)$ denotes the predicted value for sample n, and y_n represents the label for the sample.

In summary, our method P-RMF is designed with three learning objectives: $\mathcal{L}_{VAE}$, $\mathcal{L}_{DR}$, and $\mathcal{L}_{SP}$. Therefore, the total loss can be expressed as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{SP} + \lambda_2 \mathcal{L}_{VAE} + \lambda_3 \mathcal{L}_{DR} \tag{12}$$

where λ_1, λ_2 and λ_3 are weighted hyperparameters.

4 Experiments

4.1 Datasets

We performed a comprehensive experimental analysis on three widely used public datasets MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018) and SIMS (Yu et al., 2020). The details of the datasets are shown in Appendix A.

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**

Table 1: Robustness comparison of overall performance on MOSI and MOSEI under intra-modal missingness.

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
CENET	68.13	57.90	53.17	22.29	0.589	0.107
TETFN	73.58	68.67	56.91	33.42	0.505	0.387
TFR-Net	68.13	58.70	52.89	26.52	0.661	0.169
ALMT	71.85	76.21	56.47	34.16	0.509	0.372
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**

Table 2: Robustness comparison of overall performance on SIMS under intra-modal missingness.

4.2 Implementation Details

Evaluation Metrics. MOSI and MOSEI: Acc-2, Acc-5, Acc-7, F1 scores, MAE, and Corr. SIMS: Acc-2, Acc-3, Acc-5, F1 scores, MAE, and Corr. Appendix B provides detailed of the metrics.

Experimental Setup: Detailed Experimental Setup and parameters are provided in Appendix B.

4.3 Baselines

We conduct a fair comparison with several advanced and state-of-the-art methods, including **complete-modality methods:** MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), CENET (Wang et al., 2022), TETFN (Wang et al., 2023a), ALMT (Zhang et al., 2023), CubeMLP (Sun et al., 2022), and DMD (Li et al., 2023b); as well as **missing-modality methods:** TFR-Net (Yuan et al., 2021), LNLN (Zhang et al., 2024), MCTN (Pham et al., 2019), TransM (Wang et al., 2020), SMIL (Ma et al., 2021), GCNet (Lian et al., 2023), and CorrKD (Li et al., 2024b).

4.4 Robustness Comparison

We evaluate the robustness and effectiveness of P-RMF under two scenarios: Intra-modal missingness and Inter-modal missingness.

Robustness Comparison for Intra-modal Missingness. Following previous work (Zhang et al., 2024), we set the missing rate to predefined values ranging from 0 to 0.9, with an increment of 0.1, to simulate the test conditions of intra-modal random missingness. For each method, we calculate the average results at different missing rates, reflecting the model's overall performance under varying noise levels. More detailed test results are provided in Appendix D.

Tables 1 and 2 show the average robustness evaluation results of different models on the MOSI, MOSEI, and SIMS datasets under varying intra-modal missingness. As shown in these tables, P-RMF significantly outperforms existing models in noise resistance and achieves SOTA or competitive results across multiple evaluation metrics.

As shown in Table 1, compared to the SOTA missing-modality method LNLN, P-RMF improves all metrics on the MOSI and MOSEI datasets by an average of 0.36% and 2.46%, respectively. This indicates that P-RMF, with its proxy-dominated joint multimodal representation, outperforms the LNLN text-dominated approach in noise resistance. However, P-RMF achieves only suboptimal performance on metrics such as Acc-7, possibly due to an imbalance in data distribution. For example, in the MOSEI dataset, 67.85% of sentiment values fall within the range of -1 to 1, with imbalances further exacerbated in noisy scenarios. Since P-RMF focuses more on learning consistent features, the imbalanced distribution increases its difficulty in learning fine-grained tasks when data is missing.

As shown in Table 2, P-RMF achieves significant improvements in several metrics on the SIMS dataset. Compared with the suboptimal results of LNLN, P-RMF improves 1.25% and 4.28% on Acc-2 and Corr, respectively, which verifies its robustness under different noise scenarios. However, the

Dataset	Model	Testing Condition							
		{t}	{a}	{v}	{t, a}	{t, v}	{a, v}	{t, a, v}	Avg.
MOSI	Self-MM	67.80	40.95	38.52	69.81	74.97	47.12	**84.64**	60.54
	CubeMLP	64.15	38.91	43.24	63.76	65.12	47.92	84.57	58.24
	DMD	68.97	43.33	42.26	70.51	68.45	50.47	84.50	61.21
	MCTN	75.21	59.25	58.57	77.81	74.82	64.21	80.12	70.00
	TransM	77.64	63.57	56.48	82.07	80.90	67.24	82.57	72.92
	SMIL	78.26	67.69	59.67	79.82	79.15	71.24	82.85	74.10
	GCNet	80.91	65.07	59.67	**84.73**	**83.58**	70.02	83.20	75.31
	CorrKD	81.20	66.52	60.72	83.56	82.41	**73.74**	83.94	76.01
	P-RMF	**81.36**	**71.44**	**70.32**	82.10	81.94	73.11	84.37	**77.81**
MOSEI	Self-MM	71.53	43.57	37.61	75.91	74.62	49.52	83.69	62.35
	CubeMLP	67.52	39.54	32.58	71.69	70.06	48.54	83.17	59.01
	DMD	70.26	46.18	39.84	74.78	72.45	52.70	84.78	63.00
	MCTN	75.50	62.72	59.46	76.64	77.13	64.84	81.75	71.15
	TransM	77.98	63.68	58.67	80.46	78.61	62.24	81.48	71.87
	SMIL	76.57	65.96	60.57	77.68	76.24	66.87	80.74	72.09
	GCNet	80.52	66.54	61.83	81.96	81.15	69.21	82.35	74.79
	CorrKD	80.76	66.09	62.30	81.74	81.28	71.92	82.16	75.18
	P-RMF	**81.91**	**75.91**	**73.19**	**84.61**	**85.17**	**76.88**	**85.48**	**80.45**

Table 3: Performance comparison under varying inter-modal missingness conditions on MOSI and MOSEI. For example, the symbol "{t}" indicates that only the text modality is available.

F1 performance of P-RMF is inferior to that of LNLN. As can be seen from Fig. 5(c) and (f), the F1 performance of LNLN improves as the missing rate increases while the MAE performance continues to decrease. This suggests that the LNLN is biased in the face of missing data and the unbalanced data makes the model tend to perform lazy behavior in high-noise scenarios, i.e., predicting a higher proportion of categories in the training set (Zhang et al., 2024). This shows that the model struggles to learn effective predictive knowledge and instead relies on lazy to maintain accuracy.

Fig. 5 shows the performance curves of several advanced methods under different missing rates, providing an intuitive reflection of model robustness. It can be observed that as the missing rate increases, the performance of all models declines. Model structures optimized for modal missingness (e.g., TFR-Net) show a significant drop in robustness under varying levels of random data missingness. However, we proposed P-RMF consistently outperforms other models in most cases, demonstrating exceptional robustness.

Robustness Comparison for Inter-modal Missingness. Following previous work (Li et al., 2024b), we removed entire modalities from samples in the MOSI and MOSEI datasets to simulate inter-modal missingness test conditions. We evaluate model performance in different inter-modal missingness scenarios using the F1 score. Results for other metrics can be found in Appendix E.

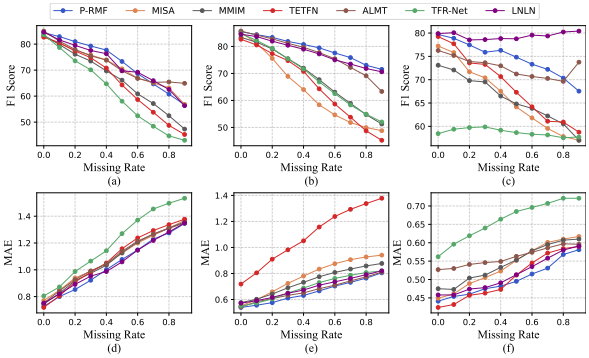

Figure 5: Performance curves of various missing rates. (a), (b) and (c) is the F1 curves on MOSI, MOSEI and SIMS. (d), (e) and (f) is the MAE curves on MOSI, MOSEI and SIMS.

The following key conclusions can be drawn from Table 3: (i) Inter-modal missingness leads to performance degradation across all models, indicating that complementary information from hetero-

Method	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
P-RMF	**71.53 / 72.81**	71.69 / **72.93**	**38.50**	**34.19**	**1.038**	0.525
w/o PMG	70.38 / 71.79	70.54 / 71.82	37.35	33.53	1.055	0.517
w/o PDDI	71.24 / 72.15	71.10 / 72.34	34.49	30.95	1.107	0.476
w/o DR	71.51 / 72.07	**71.76** / 72.14	35.85	32.10	1.066	**0.526**

Table 4: Ablation experiments on MOSI.

geneous modalities enhances sentiment semantics in joint representation; (ii) The performance of the text modality outperforms other modalities, suggesting that the text modality contains more rich knowledge information; (iii) Our proposed P-RMF exhibits significantly better noise resistance under different inter-modal missingness conditions compared to other models. Compared with the suboptimal model CorrKD, P-RMF improves the average performance on the MOSI and MOSEI datasets by 2.37% and 7.01%, respectively; (iv) Existing training models for missing data perform well in robustness tests for missing modalities, while complete modality training models perform better in complete modality tests. Our method, P-RMF, demonstrates optimal robustness in modality missing scenarios and high performance in complete modality scenarios. For example, in the complete modal test on MOSEI, P-RMF achieves an F1 score of 85.48, improving by 0.83% over the second-best method, DMD (a complete modality method).

4.5 Ablation Study

To investigate the contribution of the core modules of P-RMF, we conducted ablation experiments on MOSI, calculating the average test results across different missing rates.

As shown in Table 4, removing different modules of P-RMF leads to a decrease in overall performance. Specifically, when the PMG module is removed, binary classification tasks (such as Acc-2 and F1) show a significant decline, while multiclass tasks (such as Acc-5 and Acc-7) remain relatively stable. In contrast, removing the PDDI module has the opposite effect compared to removing PMG. This indicates that the PMG module helps to learn stable and consistent multimodal representation, while the introduction of PDDI enhances the diversity of the joint representation by injecting unimodal feature information into the proxy joint representation generated by PMG. Thus, PDDI and PMG complement each other to guarantee stability across different tasks. In addition, Fig. 6 shows

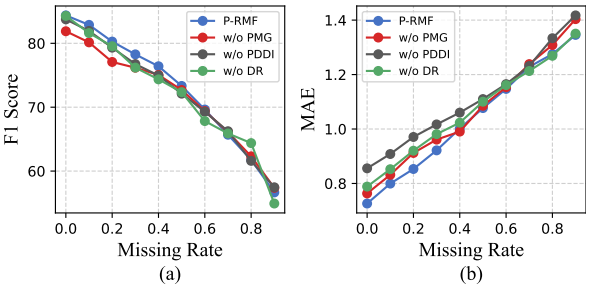

Figure 6: Ablation results of different various missing ratios on MOSI.

the experimental results of P-RMF ablation on the MOSI dataset with different missing rates.

As observed in Fig. 6, even at lower missing rates, removing the PMG module leads to a significant drop in F1 performance, indicating that the PMG module contributes to the model's stability. When the PDDI module is removed, F1 performance remains relatively stable, but the MAE value is higher, suggesting that the absence of PDDI increases prediction errors.

5 Conclusion

This paper proposes a data-centric robust multimodal sentiment analysis method, Proxy-Driven Robust Multimodal Fusion (P-RMF). P-RMF shows outstanding robustness in both uncertain missing and complete multimodal data scenarios. Specifically, P-RMF learns stable modality representation from the latent space of the data's Gaussian distribution and quantifies intrinsic modality uncertainty to obtain stable and robust multimodal joint representation (i.e., proxy modality). In the proxy-driven cross-modal injection framework, modality-specific features with varying weights is iteratively injected into the proxy modality, enhancing its diversity representation. Comprehensive experiments demonstrate that P-RMF outperforms existing models in noise resistance and robustness. P-RMF considers both the consistency of data distribution and intrinsic uncertainty, offering a novel method for robust MSA with incomplete data.

Limitations

Although P-RMF achieves performance improvements in exploring robust multimodal representation of arbitrarily uncertain missing data, there are still some limitations that need to be addressed or clarified in future work. Firstly, due to limitations in resources and experimental conditions, we did not replicate all baseline methods but instead referred to the detailed test results from (Zhang et al., 2024) and (Li et al., 2024b) in various missing data scenarios. To ensure fairness in the experiments, we used the same datasets and experimental design as in (Zhang et al., 2024) and (Li et al., 2024b), with the complete experimental setup and detailed results provided in Appendices B, D, and E. Secondly, while this paper simulates both inter-modal and intra-modal missing scenarios, real-world environments may involve more complex joint missingness, which imposes higher demands on model performance and offers opportunities for future optimization. Finally, although P-RMF achieves significant improvement in overall robustness, P-RMF does not always outperform other methods in some metrics due to the stochastic nature of the missing data noise. It is also worth investigating how to balance the performance of the model under different noise levels in the future.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62176084, and Grant 62176083, and in part by the National Key Research and Development Program of China under Grant 2023YFC3604704.

References

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

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18 – 24.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905.

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

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

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Donggeun Kim and Taesup Kim. 2024. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision*, pages 171–187. Springer.

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*, volume 38, pages 10074–10082.

Mingcheng Li, Dingkang Yang, and Lihua Zhang. 2023a. Towards robust multimodal sentiment analysis under uncertain signal missing. *IEEE Signal Processing Letters*.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024b. 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.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023b. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640.

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*, 45:8419–8432.

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*, volume 35, pages 2302–2310.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6892–6899.

Fan Qian, Jiqing Han, Yongjun He, Tieran Zheng, and Guibin Zheng. 2023. Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12966–12978, Toronto, Canada. Association for Computational Linguistics.

Upendra Singh, Kumar Abhishek, and Hiteshwar Kumar Azad. 2024. A survey of cutting-edge multimodal sentiment analysis. *ACM Computing Surveys*, 56:1–38.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3722–3729.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023a. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 658–670, Toronto, Canada. Association for Computational Linguistics.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023b. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15:309–325.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. 2023a. 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. 2022. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921.

Yuanzhi Wang, Zhen Cui, and Yong Li. 2023b. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22034.

Yuanzhi Wang, Zhen Cui, and Yong Li. 2023c. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22034.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of the web conference 2020*, pages 2514–2520.

Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:12113–12132.

Chuanpeng Yang, Fuqing Zhu, Yaxin Liu, Jizhong Han, and Songlin Hu. 2024. Uncertainty-aware cross-modal alignment for hate speech detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16973–16983, Torino, Italia. ELRA and ICCL.

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, Online. Association for Computational Linguistics.

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*, volume 35, pages 10790–10797.

Yakun Yu, Mingjun Zhao, Shi-ang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. 2023. ConKI: Contrastive knowledge injection for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13610–13624, Toronto, Canada. Association for Computational Linguistics.

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*.

Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1545–1554.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. 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.

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

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Fuji Ren. 2024. Kebr: Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 5732–5741.

A Datasets

MOSI. The MOSI dataset consists of speakers expressing their opinions on topics such as movies across 93 YouTube videos, containing 2,199 video clips with sentiment annotations. The sentiment intensity of each clip is labeled on a scale from -3 (strongly negative) to +3 (strongly positive).

MOSEI. The MOSEI dataset is an extended version of MOSI, covering 250 different topics and containing 22,856 annotated video clips. The sentiment annotation method is the same as that of MOSI.

SIMS. The SIMS dataset consists of 2,281 video clips. Each sample has one multimodal sentiment label and three unimodal sentiment labels, with sentiment scores ranging from -1 (negative) to +1 (positive).

The statistics of these datasets are summarized in Table 5.

B Implementation Details

Evaluation Metrics: For the MOSI and MOSEI datasets, we report binary classification accuracy (Acc-2), five-class classification accuracy (Acc-5), seven-class classification accuracy (Acc-7), as well as F1 scores and mean absolute error (MAE) associated with Acc-2. The Acc-2 and F1 scores are reported in two forms, using the split marker "-/-": the first score represents negative/non-negative (including 0), and the second score represents negative/positive. For the SIMS dataset, we report Acc-2, Acc-3, Acc-5, F1 scores, MAE, and correlation (Corr). Acc-2, Acc-3, Acc-5, and Acc-7 represent the percentage of correct predictions within their respective sentiment intervals. For example, Acc-7 indicates the accuracy across seven sentiment intervals from -3 to +3. Except for MAE, higher values for all metrics indicate better model performance.

Experimental Setup: All models are trained using the PyTorch framework on an NVIDIA RTX A40 with 60GB of memory. For each dataset, training is performed with different random seeds (1111, 1112, and 1113), and results are averaged. The Adam optimizer is used, with BERT as the backbone. The learning rate is set to 1e-4, the batch size to 32, and training lasts for 100 epochs. During each epoch, 50% of the training samples are randomly selected, and for each modality, a random percentage of information (from 0% to 100%) is erased to simulate data missingness. Detailed parameters are provided in Table 6.

C Efficiency Analysis

Table 7 shows the computational overhead of the proposed P-RMF model on the MOSI dataset (Zadeh et al., 2016), and compares it with the state-of-the-art model LNLN (Zhang et al., 2024) for incomplete modality under the same experimental setup. Compared to LNLN, P-RMF achieves a shorter runtime per epoch while maintaining a similar number of parameters, indicating that its model structure is more optimized and computationally efficient.

D Robustness Evaluation of P-RMF under Intra-modal Missingness

Following previous work (Zhang et al., 2024), we set the missing rate to predefined values ranging from 0 to 0.9, with an increment of 0.1, to simulate the test conditions of intra-modal random missingness. Like previous work(Zhang et al., 2024), we did not evaluate at $r = 1.0$, as this would imply complete data erasure from each modality, rendering the experiment non-informative.

Tables 8, 9, and 10 present the detailed results of the robustness tests for our model, P-RMF, on the MOSI, MOSEI, and SIMS datasets under various data missing rates. It can be observed that the model performance decreases continuously with the increase of missing rate. This is primarily due to the increase in missing modality rates, which not only reduces the available information for the model but also introduces indirect issues such as distribution shifts, weakened feature correlations, and amplified noise. The combined impact of these factors leads to a continuous decline in model performance, making it challenging for the model to maintain stable, state-of-the-art results.

E Robustness Evaluation of P-RMF under Inter-modal Missingness

Following previous work (Li et al., 2024b), we removed entire modalities from samples in the MOSI and MOSEI datasets to simulate inter-modal missingness test conditions.

Tables 11, 12, and 13 present the robustness evaluation results of our model, P-RMF, on the MOSI, MOSEI, and SIMS datasets under inter-modal missing conditions. It can be observed that model performance generally improves with the increase in modality information. The performance of the text modality typically surpasses that of other unimodal, indicating that the text modality contains

Dataset	Speaker	Video clip	Train	Valid	Test	Language
MOSI	93	2199	1284	229	686	English
MOSEI	1000	22856	16326	1871	4659	English
SIMS	474	2281	1368	456	457	Chinese

Table 5: Dataset statistics.

Hyper-parameter	MOSI	MOSEI	SIMS
d_t	768	768	768
d_a	5	74	33
d_v	20	35	709
L	4	4	4
λ_1, λ_2, λ_3	1,0.5,0.1	1,0.5,0.1	1,0.5,0.1
Batch size	32	32	32
Epoch	100	100	100
Optimizer	Adam	Adam	Adam
Vector Length T	8	8	39
Learning rate	1e-4	1e-4	1e-4
Fully connected layer	128	128	128

Table 6: Hyper-parameters setting.

Model	Parameters	Time / Epoch
P-RMF	117 M	18 s
LNLN	116 M	24 s

Table 7: Computational overhead.

richer knowledge information. It is worth noting that in all test scenarios across the datasets, even though the Acc-2 or F1 scores improve, the model's other performance metrics significantly drop whenever the text modality is missing. For example, in Table 13, under testing conditions such as {a}, {v}, or {a,v}, although the F1 score is higher, metrics like Acc-5, Acc-7, MAE, and Corr see a notable decline. This indicates that missing the text modality causes the model to lose crucial semantic information, while the sentiment cues provided by the audio and visual modalities are relatively weak and cannot compensate for the loss of the text modality. Therefore, even with a higher F1 score in binary classification tasks, due to the loss of key features, the model's performance in more complex multi-classification and regression tasks still significantly declines.

F Independent ablation on MOSI.

Independent ablation of the three loss functions in the Proxy Modality Generation (PMG) module, As shown in Table 14. SA, CMA, and SR represent Standardized Alignment, Cross-Modal Alignment, and Sample Reconstruction within the PMG module.

G Hyperparameter sensitivity analysis.

The model has three λ_1, λ_2, λ_3 weight hyperparameters. Preliminary tests showed that it was difficult to find a set of hyperparameters that performed optimally across all datasets. Based on prior experience [1], we set the main task weight λ_1 to 1 and the reconstruction loss weight λ_3 to 0.1. We mainly analyzed the impact of different weights λ_2 for constraint losses in Proxy Modality Generation. The detailed analysis is shown in Table 15.

Missing Rate r	MOSI					
	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
0.0	82.65 / 84 15	82.69 / 84 37	48 83	44 31	0.726	0.782
0.1	81.34 / 82 62	81.35 / 82 89	47 52	42 13	0.800	0.730
0.2	78.13 / 79 57	78.11 / 80 97	44 75	40 38	0.853	0.668
0.3	75.80 / 76 83	75.82 / 79 27	42 71	39 21	0.922	0.621
0.4	73.76 / 75 46	74.09 / 77 71	40 67	35 59	1.001	0.584
0.5	71.28 / 73 02	71.66 / 73 33	37 90	33 67	1.077	0.523
0.6	67.35 / 68 75	67.64 / 68 64	33 24	29 30	1.147	0.432
0.7	65.16 / 66 16	65.33 / 64 69	32 94	27 84	1.229	0.383
0.8	61.08 / 62 04	61.22 / 60 76	29 74	25 97	1.275	0.316
0.9	58.75 / 59 45	59.01 / 56 66	26 68	23 49	1.346	0.212
Avg.	71.53 / 72.81	71.69 / 72.93	38.50	34.19	1.038	0.525

Table 8: Robustness evaluation results of P-RMF on MOSI under various rates of intra-modal missing data.

Missing Rate r	MOSEI					
	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
0.0	83.62 / 85.20	83 68 / 85.48	52.09	49 77	0.539	0.767
0.1	82.79 / 83.98	82 94 / 84.37	51.45	49 04	0.556	0.748
0.2	82.25 / 82.97	82 58 / 83.37	49.35	47 91	0.576	0.722
0.3	80.88 / 81 26	81 40 / 81.86	47.78	45 95	0.611	0.683
0.4	79.76 / 79 97	80 58 / 80.74	46.73	45 59	0.631	0.653
0.5	78.64 / 78 62	79 74 / 79.49	44.90	43 94	0.666	0.601
0.6	77.44 / 76 11	78 97 / 77.58	43.12	42 26	0.703	0.545
0.7	75.87 / 74 64	78 12 / 75.88	42.41	41 73	0.733	0.481
0.8	74.46 / 70 75	77 91 / 73.03	41.00	40 46	0.764	0.401
0.9	72.59 / 67 86	77 95 / 71.51	39.90	39 62	0.805	0.289
Avg.	78.83 / 78.14	80.39 / 79.33	45.87	44.63	0.658	0.589

Table 9: Robustness evaluation results of P-RMF on MOSEI under various rates of intra-modal missing data.

Missing Rate r	SIMS					
	Acc-2	F1	Acc-3	Acc-5	MAE	Corr
0.0	78.34	79.69	60.61	38.95	0.441	0.550
0.1	77.24	78.88	59.52	37.72	0.454	0.530
0.2	76.23	77.46	59.30	37.05	0.460	0.512
0.3	75.66	75.89	56.89	36.89	0.475	0.483
0.4	75.32	76.30	55.8	36.54	0.482	0.472
0.5	73.61	74.83	54.83	34.79	0.495	0.404
0.6	72.12	73.32	53.05	33.92	0.515	0.383
0.7	71.58	72.22	52.52	32.23	0.531	0.369
0.8	69.51	70.35	49.89	31.07	0.568	0.269
0.9	66.77	67.54	45.08	29.10	0.581	0.168
Avg.	73.64	74.65	54.75	34.83	0.500	0.414

Table 10: Robustness evaluation results of P-RMF on SIMS under various rates of intra-modal missing data.

Testing Condition	MOSI					
	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
{l}	80.90 / 81.01	80.89 / 81.36	47.46	42.57	0.777	0.759
{a}	55.03 / 56.85	71.72 / 71.44	20.99	20.55	1.367	0.107
{v}	54.96 / 56.01	70.72 / 70.32	20.26	19.83	1.368	0.099
{l, a}	81.05 / 82.16	81.03 / 82.10	48.69	43.29	0.776	0.760
{l, v}	80.90 / 82.01	80.89 / 81.94	48.40	43.15	0.777	0.759
{a, v}	55.95 / 58.42	71.98 / 73.11	20.85	20.41	1.366	0.109
{l, a, v}	82.65 / 84.15	82.69 / 84.37	48.83	44.31	0.726	0.782
Avg.	70.21 / 71.52	77.13 / 77.81	36.50	33.44	1.022	0.482

Table 11: Robustness evaluation results of P-RMF under inte-modality missingness on the MOSI dataset.

Testing Condition	MOSEI					
	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
{l}	81.52 / 82.26	82.11 / 81.91	50.46	49.09	0.562	0.757
{a}	71.02 / 62.85	83.06 / 75.91	41.74	41.25	0.838	0.115
{v}	70.52 / 61.77	82.14 / 73.19	33.68	33.59	0.828	0.209
{l, a}	83.58 / 84.62	83.60 / 84.61	50.42	49.04	0.565	0.757
{l, v}	83.52 / 85.20	83.49 / 85.17	51.87	49.41	0.559	0.765
{a, v}	70.68 / 63.34	82.16 / 76.88	37.63	33.59	0.822	0.211
{l, a, v}	83.62 / 85.20	83.68 / 85.48	52.09	49.77	0.539	0.767
Avg.	77.78 / 75.03	82.89 / 80.45	45.41	43.68	0.673	0.512

Table 12: Robustness evaluation results of P-RMF under inte-modal missingness on the MOSEI dataset.

Testing Condition	SIMS					
	Acc-2	F1	Acc-3	Acc-5	MAE	Corr
{l}	77.90	79.37	60.34	38.29	0.441	0.550
{a}	69.37	81.91	54.27	21.66	0.640	0.052
{v}	69.37	81.91	54.27	21.44	0.639	-0.013
{l, a}	77.02	78.22	60.61	38.29	0.440	0.537
{l, v}	78.12	79.87	60.39	38.29	0.442	0.537
{a, v}	69.37	81.91	54.27	21.44	0.636	0.021
{l, a, v}	78.34	79.69	60.61	38.95	0.441	0.550
Avg.	74.21	80.41	57.82	31.19	0.526	0.319

Table 13: Robustness evaluation results of P-RMF under inte-modal missingness on the SIMS dataset.

Method	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
P-RMF	**71.53 / 72.81**	**71.69 / 72.93**	**38.50**	**34.19**	1.038	**0.525**
w/o PMG	70.38 / 71.79	70.54 / 71.82	37.35	33.53	1.055	0.517
w/o SA	70.84 / 72.11	70.98 / 72.16	37.88	33.87	1.134	0.516
w/o CMA	70.77 / 71.79	70.54 / 71.82	37.76	33.53	1.120	0.508
w/o SR	71.35 / 71.55	71.14 / 71.46	38.17	33.80	**1.033**	0.520

Table 14: Independent ablation on MOSI.

λ_2	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
0.1	71.20 / 72.44	71.31 / 72.65	37.97	34.15	1.055	0.500
0.2	71.43 / 72.65	71.50 / 72.77	38.31	33.88	1.053	0.507
0.3	71.50 / 72.76	71.63 / 72.78	38.33	34.11	1.047	0.515
0.4	71.42 / 72.73	71.34 / 72.51	**38.62**	34.02	1.039	0.522
0.5	71.53 / **72.81**	**71.69 / 72.93**	38.50	**34.19**	1.038	0.525
0.6	71.51 / 72.80	71.57 / 72.88	38.55	34.11	**1.033**	**0.526**
0.7	**71.55** / 72.77	71.66 / 72.63	37.99	33.75	1.040	0.516
0.8	71.44 / 72.61	71.22 / 72.56	38.33	34.01	1.051	0.514
0.9	71.32 / 72.53	71.27 / 72.44	37.66	33.55	1.063	0.505
1.0	71.43 / 74.56	71.33 / 72.62	37.87	33.76	1.056	0.508

Table 15: Hyperparameter sensitivity analysis.