MissMAC-Bench: Building Solid Benchmark for Missing Modality Issue in Robust Multimodal Affective Computing

Ronghao Lin[1,2], **Honghao Lu**[2], **Ruixing Wu**[2], **Aolin Xiong**[2], **Qinggong Chu**[2], **Qiaolin He**[2], **Sijie Mai**[3], **Haifeng Hu**[*2,4]

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]School of Electronics and Information Technology, Sun Yat-Sen University
[3]School of Computer Science, South China Normal University
[4]Pazhou Laboratory
linrh@szu.edu.cn; huhaif@mail.sysu.edu.cn

Abstract

Current Multimodal Affective Computing (MAC) systems heavily rely on the completeness of multiple modalities to accurately understand human's affective state. However, in real-world scenarios, the availability of modality data is often dynamic and uncertain, leading to substantial performance fluctuations due to the distribution shifts and semantic deficiencies of the incomplete multimodal inputs. Known as the missing modality issue, this challenge poses a critical barrier to the robustness and practical deployment of MAC models. To systematically quantify this issue, we introduce **MissMAC-Bench**, a comprehensive benchmark designed to establish fair and unified evaluation standards from the perspective of cross-modal synergy. Two guiding principles are proposed, including no missing prior during training, and one single model capable of handling both complete and incomplete modality scenarios, thereby ensuring better generalization. Moreover, to bridge the gap between academic research and real-world applications, our benchmark integrates evaluation protocols with both fixed and random missing patterns at the dataset and instance levels. Extensive experiments conducted on 3 widely-used language models across 4 datasets validate the effectiveness of diverse MAC approaches in tackling the missing modality issue. Our benchmark provides a solid foundation for advancing robust MAC and promotes the development of multimedia data mining. Our code is released in https://github.com/RH-Lin/MissMAC-Bench.

1 Introduction

> Integration of information from multiple sensory channels is crucial for understanding tendencies and reactions in humans.
>
> —— Partan and Marler, 1999

With the rapid growth of online platforms and social media, vast amounts of multimodal data capturing diverse forms of human interaction and communication are now widely available on the web. These data encompass synchronized streams of language, speech, and visual cues, enabling the discovery of affective patterns, behavioral insights, and contextual information embedded within human communication. (Baltrušaitis et al., 2019). Emerged as a promising research area, Multimodal affective computing (MAC) aims to integrate heterogeneous signals from text, audio and vision modalities in speaking videos and predict human's affective states, such as sentiment scores and emotion classes (Poria et al., 2020). As a subfield of multimedia content analysis, previous MAC models have primarily devoted to learning productive fusion-based multimodal representations by designing delicate multimodal fusion techniques (Poria et al., 2017; Zhao et al., 2024). Given its critical role in real-world applications including digital marketing, distance education, and human-computer interaction (Liang et al., 2024), MAC models are expected to exhibit strong generalization and robust performance in downstream inference to enable broader applicability.

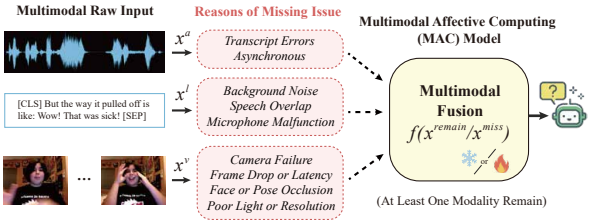


Figure 1: Missing modality issue for robust multimodal affective computing in downstream application.

Most MAC approaches assume that all predefined modalities are present in complete pairs during inference to reach superior performance. However, in downstream applications, the completeness of modalities can not be guaranteed due to numerous reasons including sensor failure or environment interference (Wu et al., 2024; Teng et al., 2024), as shown in Figure 1. The lack of modalities leads to

[*] Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 20669–20684
October 24-29, 2026 ©2026 Association for Computational Linguistics

significant performance variation and degradation when applying pre-trained models in the real world, named as missing modality issue (Ma et al., 2022; Liang et al., 2024; Zhao et al., 2025).

Although previous methods have partially mitigated the impact of missing modality(Wu et al., 2026), there remains a lack of unified evaluation standards. Inconsistent evaluation settings and metrics make it challenging to fairly compare different methods across diverse scenarios (Lin and Hu, 2023a; Lian et al., 2023). Moreover, prior studies have not clearly distinguished intra- and inter-modal robustness (Details in Appendix A) during evaluation, limiting the understanding of multimodal model behavior under various missing conditions (Wu et al., 2024).

In addition, most methods are evaluated only under specific and method-dependent settings, failing to comprehensively cover the holistic spectrum of diverse missing modality scenarios with sufficient flexibility and efficiency (Ma et al., 2021). Lastly, to enhance the systematical evaluation on model's generalization ability, different MAC subtasks and datasets must be carefully considered when addressing the missing modality issue, as MAC requires rich semantic understanding, with diverse affective contributions from each modality (Ma et al., 2022; Lin and Hu, 2023b).

To ensure fair comparison, establishing standardized evaluation benchmark is both necessary and critical for advancing research in missing modality issue of MAC. Therefore, focusing on revealing cross-modal synergy, we present a solid benchmark, named **MissMAC-Bench**, to evaluate models' inter-modal robustness in MAC task. Our benchmark introduces two main principles to conduct practical robustness assessment and employs both fixed and random missing protocols, with random evaluation conducted at dataset- and instance-level to comprehensively cover diverse missing scenarios. We further provide analysis on 27 current methods to consistently show the applicability ability of different types of methods across multimodal sentiment analysis and emotion recognition subtasks. Our contribution can be summarized as:

- **Unified evaluation standards**: Two principles are introduced to solidly simulate the performance of pre-trained models in practical applications, including no missing prior and single model to handle various incomplete scenarios;
- **Comprehensive missing modality protocols**: Both fixed and random evaluation protocols are adopted, with random evaluation conducted at both dataset and instance level to bridge the gap between academia and real-world application.
- **Extensive empirical validation**: To facilitate a systematic assessment of generalization capabilities, the benchmark is evaluated across 2 subtasks, 4 datasets, and 3 widely-used language models, providing a thorough evaluation of cross-modal synergy and inter-modal robustness.

2 Method Taxonomy

We categorize previous methods in addressing the missing modality issue into four types of techniques described as follows.

2.1 Data Augmentation Methods

According to task goal, Multimodal affective computing (MAC) can be divided into sentiment score regression and emotion category classification (Poria et al., 2017; A.V. et al., 2024). Previous MAC models have primarily devoted to developing diverse multimodal fusion methods, including feature subspace disentanglement (Hazarika et al., 2020), attention mechanism (Rahman et al., 2020), knowledge guidance (Yu et al., 2021; Lin and Hu, 2023b), information theory (Han et al., 2021b; Xiao et al., 2024), sequence-based recurrent network (Majumder et al., 2019; Hu et al., 2021) and graph neural network (Joshi et al., 2022; Nguyen et al., 2023). When these models encounter missing modality during downstream inference, their performance drops significantly (Hazarika et al., 2022; Lin and Hu, 2023a). A straightforward approach is to apply general data augmentation techniques to enhance robustness while preserve the fusion networks. For instance, Hazarika et al. (2022) introduce dropping and noise into the language modality while Lin and Hu (2024) apply mixup strategy at both data and feature levels to create augmented data. However, these methods often degrade the original fusion performance, highlighting the need for more effective techniques to balance robustness on incomplete multimodal learning and the pre-trained complete capability.

2.2 Alignment-based Methods

Considering input with incomplete and complete modalities as different views of the same multimodal samples, previous methods conduct multi-view learning (Zhang et al., 2022; Yu et al., 2025)

to align the incomplete and complete representations without imputation. Techniques including contrastive learning (Poklukar et al., 2022; Lin and Hu, 2023a), canonical correlation analysis (Andrew et al., 2013; Wang et al., 2015; Sun et al., 2020), adversarial learning (Yuan et al., 2023) and optimal transport (Han et al., 2022) have been explored for this purpose. These alignment-based methods, focus on exploiting the commonly shared and invariant features to compensate missing information. Although they can effectively improve the consistency between complete and incomplete multimodal representations, their performance remains suboptimal due to the huge modality gap in multimodal fusion (Liang et al., 2022).

2.3 Generation-based Methods

Another promising approach is leveraging generative models to synthesize missing modality data, and then using the generated information to supplement the incomplete inputs. Current generation-based methods utilize current generative models include auto-encoders (Tran et al., 2017; Pham et al., 2019; Zhao et al., 2021), variational auto-encoders (Wu and Goodman, 2018; Shi et al., 2019; Vasco et al., 2022; Cheng et al., 2025), graph completion networks (Lian et al., 2023), and diffusion models (Wang et al., 2023) to translate information across different modalities or generate the missing information. These generative methods effectively reconstruct the missing information due to their powerful imputation capabilities. Meanwhile, they can mostly maintain the original multimodal fusion performance when no modifications are required for the original fusion networks. However, these methods demand higher computation costs and complex hierarchical networks, making them less practical for real-time applications in real world.

2.4 Adaption-based Methods

With the emergence of multimodal large language model (Zhang et al., 2024a), leveraging additional parameter-efficient fine-tuning techniques to adapt the original multimodal model into various missing scenarios have also been increasingly explored, such as prompt tuning (Lee et al., 2023; Guo et al., 2024), mixture-of-expert networks (Xu et al., 2024; Han et al., 2024; Gao et al., 2024), and adapter module (Zhang et al., 2024b; Zhu et al., 2025). Besides, since there exists optimization conflict between complete and incomplete multimodal learning and imbalance contribution of diverse modalities, some methods adopt adaptive gradient manipulation (Wei et al., 2025; Lin and Hu, 2024; Wang et al., 2024) to adjust the contribution of each modality in various missing scenarios. Since the adaptation-based methods can adjust existing frameworks with minimal additional parameters or address gradient conflicts, they generally exhibit strong generalization ability on diverse pre-trained models. Thus, an increasing number of recent researches tend to adopt such approaches to enhance the adaptability across various input scenarios.

3 Task Formulation

MissMAC-Bench aims to evaluate the robustness of Multimodal Affective Computing (MAC) models when handling incomplete inputs with missing modality, while still harnessing the potential of multimodal fusion when complete data is available. To simulate the downstream scenarios in the real world, we establish a standardized evaluation setting for both training and inference stages, guided by two main principle: (1) **No missing prior**; (2) **One unified model to handle both complete and incomplete input with various missing scenarios**.

We consider the missing modality issue in MAC with three modalities: textual, acoustic, and visual modality, denoted as $u \in \{l, a, v\}$. Given a multimodal data sample x_i as input, the complete observation is defined as $x_M \equiv \{x_i^l, x_i^a, x_i^v\} \in \mathbb{R}^{d_u}$ including language, audio and vision data, respectively, where where d_u is the dimension of the input vector. When some modalities are missing, the incomplete observation is defined as $x_m \equiv \{x_i^{remain}, x_i^{miss}\}$ where x_i^{miss} denotes the missing modality and x_i^{remain} denotes the remained ones.

During training, both x_i^{miss} and x_i^{remain} are allowed to be utilized for fully exploiting the available multimodal data. Current MAC models have presented sophisticated multimodal fusion modules to integrate information across modalities (Poria et al., 2017), providing an excellent initialization model for handling missing modality. With the pre-trained weights of multimodal fusion modules, the MAC model can be optimized either jointly or separately for both complete and incomplete inputs.

To claim models' generalizability, the first principle is to deliberately avoid introducing any prior assumptions about the missing modality in the training stage. This is crucial because the missing scenarios in real-world downstream applications are uncertain and can vary widely. Note that most pre-

vious methods (Lian et al., 2023; Wang et al., 2023; Zhang et al., 2024b; Guo et al., 2024) use the same missing rates or masks to simulate missing scenarios of data in both training and inference stage to maintain higher performance. Nevertheless, we argue that such setting will obtain a prior bias for the missing data to achieve better performance, due to the unknown missing circumstances in the real-world downstream inference (Lin and Hu, 2023a). Besides, simulating missing data during training will cause overfitting on the observed data in each iterations, leading to low training data utilization and unstable loss(Hazarika et al., 2022; Ma et al., 2022; Lin and Hu, 2024). Therefore, the circumstance of missing modality remains uncertain and unknowable during training in our benchmark.

During inference, the MAC model can operate either in complete multimodal learning with all modalities for samples x_M or in incomplete multimodal learning with one or more missing modalities. The second principle is to train each model only once and use the same pre-trained weights for evaluation across diverse modality-presence scenarios. This setup further simulates real-world scenarios, where the model cannot be modified regardless of the type or completeness of the input data.

For incomplete multimodal learning with missing modality, the x_i^{miss} part is masked out with various missing scenarios during inference. Relying solely on the remained modalities x_i^{remain}, the MAC model is expected to approach the upper bound performance referring to inference with complete modalities.

For complete and incomplete modalities input x_M and x_m, the MAC model conduct multimodal fusion $f(\cdot)$ and project the final representations into the affection space to achieve the predicted labels $\hat{y}$, formulated as follows:

$$\hat{y} = \begin{cases} f(x_M) = f(x^l, x^a, x^v) \\ f(x_m) = f(x^{remain}/x^{miss}) \end{cases} \in \mathbb{R}^1 \quad (1)$$

Additionally, the developed MAC model could be deployed for both multimodal sentiment analysis and emotion recognition subtasks, where $\hat{y}$ refers to sentiment score and emotion class, respectively. As the regression and classification subtasks of MAC, the task optimization objectives of MAC can be computed with the ground truth labels y_i and the affection prediction $\hat{y}_i$:

$$\mathcal{L}_{task} \begin{cases} \underset{regress}{=} \underset{i \in \mathcal{N}}{\mathbb{E}} \|y_i - \hat{y}_i\| \text{ or } \underset{i \in \mathcal{N}}{\mathbb{E}} \|y_i - \hat{y}_i\|_2^2 \\ \underset{classify}{=} \underset{i \in \mathcal{N}}{\mathbb{E}} \; y_i \log \hat{y}_i \end{cases} \quad (2)$$

4 Evaluation with Missing Modality Issue

Summarized from previous works (Zhang et al., 2022; Lin and Hu, 2023a; Wang et al., 2023), we aim to standardize the evaluation settings for the missing modality issue in MAC by setting two protocols.

4.1 Fixed Missing Protocol

For the input data in fixed missing protocol, one specific modality (e.g., $\{l, a\}/\{l, v\}/\{a, v\}$) or two modalities (e.g., $\{l\}/\{v\}/\{a\}$) are consistently removed across all iterations, which ensures the same missing patterns being applied either to the entire dataset or to individual instances. The goal of this protocol is to assess the contribution of each modality and bi-modal interactions.

4.2 Random Missing Protocol

As shown in Figure 2, the random missing protocol evaluates robustness in terms of both efficiency and flexibility through two evaluation levels.

- **Dataset-level Evaluation**

  We utilize Missing Rate MR to denote the proportion of missing modality relative to the total number of modalities across the entire dataset, formulated as follows:

  $$MR = \frac{\sum_{i=1}^{N} k_i}{N \times U} \leq \frac{U-1}{U}$$

  where N denotes total number of samples in datasets, $U = 3$ denotes the number of modalities in MAC, $0 \leq k_i \leq 2$ denotes the number of missing modalities for sample x_i. Thus, we set the value of missing rate $MR \in [0.1, 0.2, ..., 0.6, 0.7]$ where 0.7 denotes the most severe scenario with $\frac{2}{3}$ modalities missing.

- **Instance-level Evaluation**

  We adopt Missing Probability MP to represent the probability of each modality being missing within a single multimodal instance. Given one instance sample x_i, the probability of missing

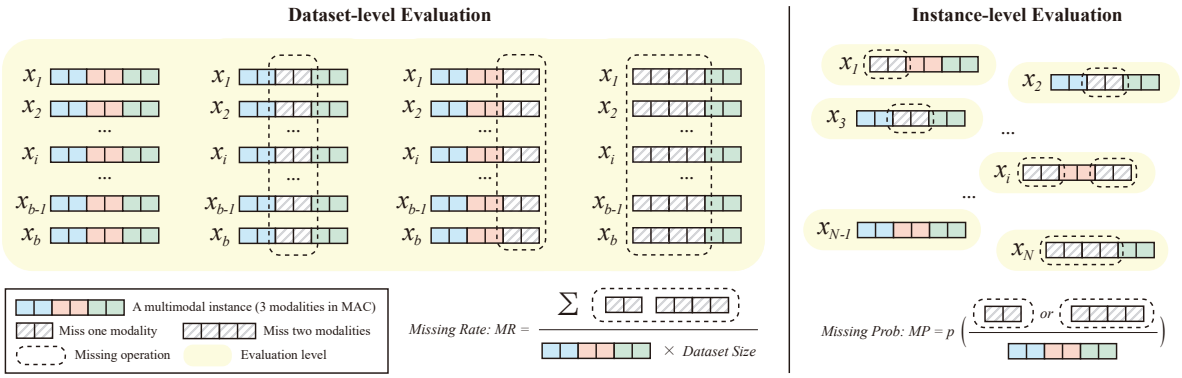


Figure 2: Illustration of random missing protocol, including dataset-level and instance-level evaluation.

exactly k_i modalities follows a Binomial Distribution, denoted as:

$$p(k_i) = \binom{U}{k_i} MP^{k_i}(1-MP)^{U-k_i}$$

With $U = 3$ and $0 \leq k_i \leq 2$, we normalize the probabilities to guarantee $p(k_i = 0) + p(k_i = 1) + p(k_i = 2) = 1$ and set $MP \in [0.1, 0.3, 0.5, 0.7, 0.9]$. Note that $MP = 1.0$ denotes the case where each instance has at least one missing modality.

To reflect real-world application scenarios, each sample has at least one available modality no matter in which evaluation settings and $MR/MP = 0.0$ denotes complete multimodal learning. Due to the space limitation, except for complete modalities input, we only report the performance with $u \in \{l\}/\{a, v\}$ as the dominant and inferior modalities for the fixed missing protocol and with $MR = 0.7/MP = 1.0$ as the most severe absent situation for the random missing protocol in Section 5. More numerical results in various missing circumstances can be found in Appendix D.

4.3 Evaluation on $C\&R$ Dimension

Since there are multiple missing circumstances no matter for dataset-level or instance-level evaluation, it's hard and dazzling to compare the performance of different models from one MR/MP to another. To better compare the robustness of different methods on random missing protocol as in the real world, we visualize the overall performance and stability of each method as an ellipse as illustrated in Fig. 3 for two dataset on each subtask, along with two complementary evaluation dimensions named as $C\&R$ dimensions:

- **Competence**: This dimension measures the model's general effectiveness by computing the average performance across all missing modality scenarios in specific missing protocol, providing a holistic view of its ability in handling incomplete inputs. It is denoted as the center of the performance ellipse.

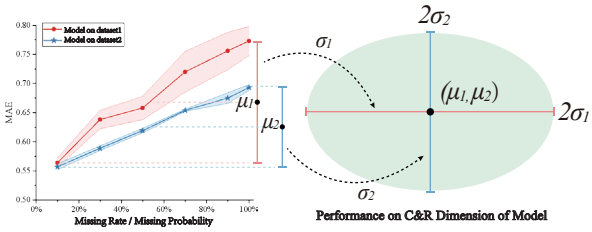


Figure 3: Illustration of $C\&R$ dimensions, where μ and σ denotes the mean and variance of the performance during various missing circumstances.

- **Resilience**: This dimension captures the performance variation across different missing modality scenarios, ranging from the least to the most severe missing conditions. It is quantified as the standard deviation of performance, reflecting how sensitive a model is to varying degrees of missing modalities. It is denoted as the axis of the performance ellipse.

5 Experiment

5.1 Sub-tasks and Datasets

We employ two public MAC sub-tasks in the proposed benchmark depicted as follows.

Multimodal Sentiment Analysis (MSA): *CMU-MOSI* (Zadeh et al., 2016) consists of 2,199 monologue utterances from 93 opinion-based YouTube videos by 89 movie reviewers, with sentiment annotations on a continuous scale from -3 (strongly negative) to $+3$ (strongly positive). *CMU-MOSEI* (Zadeh et al., 2018) expands this sentiment datasets by providing about 66 hours video clips from 250 diverse topics and 1,000 unique speakers, with each clip similarly labeled on $[-3, +3]$ Likert scale.

Table 1: Performance comparison on fixed missing protocol with input modalities set as $u \in \{l\}/\{a, v\}$.

Language Model		BERT-base / sBERT-base												DeBERTaV3-large											
Fix	Models	CMU-MOSI				CMU-MOSEI				IEMOCAP		MELD		CMU-MOSI				CMU-MOSEI				IEMOCAP		MELD	
		Acc7↑	F1↑	MAE↓	Corr↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑	Acc↑	wF1↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑	Acc↑	wF1↑
$\{l\}$	BBFN^r	44.2	84.5	0.745	0.790	51.9	85.1	0.557	0.749	51.2	52.0	59.6	58.2	-	-	-	-	-	-	-	-	-	-	-	-
	MTMD^r	46.7	83.5	0.764	0.782	53.1	85.2	0.544	0.753	58.0	57.5	57.7	58.0	-	-	-	-	-	-	-	-	-	-	-	-
	ALMT^r	48.3	85.2	0.691	0.807	53.1	82.2	0.533	0.764	53.6	53.6	59.9	58.1	50.7	89.5	0.576	0.883	51.6	83.0	0.551	0.777	48.8	45.5	59.2	57.2
	ITHP^r	43.8	82.6	0.762	0.784	52.1	84.7	0.551	0.756	56.5	56.2	56.3	56.3	49.2	87.6	0.592	0.860	50.3	83.5	0.542	0.786	50.2	49.3	57.2	56.3
	M3Net^r	-	-	-	-	-	-	-	-	47.3	45.5	63.4	60.4	-	-	-	-	-	-	-	-	-	-	-	-
	CCA	26.4	74.7	1.098	0.567	44.6	80.2	0.677	0.621	56.1	53.9	49.7	43.1	24.5	73.9	1.145	0.552	43.6	77.1	0.727	0.536	47.6	44.5	49.2	38.9
	DCCA	23.3	73.5	1.196	0.520	41.5	68.0	0.810	0.377	51.7	50.7	48.9	35.3	15.0	32.1	1.466	0.051	41.4	56.0	0.840	0.308	27.7	18.2	48.1	31.3
	DCCAE	25.8	66.4	1.278	0.508	41.6	67.9	0.806	0.336	50.1	49.2	47.7	32.0	17.9	50.2	1.358	0.179	41.2	62.8	0.804	0.304	30.7	26.5	48.1	31.3
	CPM-Net	17.2	63.9	1.335	0.345	40.3	74.7	0.832	0.230	50.4	50.5	42.0	34.8	15.5	40.0	1.423	0.105	41.4	48.6	0.839	0.105	23.4	14.1	35.0	30.8
	MM-Align	36.4	82.1	0.903	0.733	46.3	82.8	0.662	0.726	52.3	49.9	40.0	36.4	-	-	-	-	-	-	-	-	-	-	-	-
	NIAT	38.8	82.8	0.819	0.746	54.0	85.5	0.529	0.769	62.8	62.7	59.4	60.7	-	-	-	-	-	-	-	-	-	-	-	-
	MissModal	44.9	84.4	0.704	0.801	52.2	85.2	0.546	0.762	62.3	62.3	57.4	57.3	-	-	-	-	-	-	-	-	-	-	-	-
	CRA	36.3	82.8	0.900	0.749	49.3	84.9	0.576	0.741	53.8	51.3	55.3	51.6	46.2	87.3	0.609	0.863	54.4	87.4	0.504	0.806	55.1	52.6	61.9	60.2
	MVAE	43.0	83.8	0.731	0.786	51.7	83.8	0.565	0.761	56.3	55.8	54.6	54.8	-	-	-	-	-	-	-	-	-	-	-	-
	MMVAE	38.6	84.0	0.784	0.790	50.9	84.7	0.562	0.751	55.2	53.5	57.1	57.5	-	-	-	-	-	-	-	-	-	-	-	-
	MCTN	41.5	83.7	0.777	0.765	49.6	81.8	0.594	0.715	58.4	57.8	56.1	52.3	48.0	87.9	0.609	0.859	53.4	86.9	0.514	0.798	60.6	59.2	34.2	29.9
	MMIN	42.3	85.0	0.743	0.791	52.8	84.6	0.543	0.761	57.0	57.5	59.2	54.9	45.5	88.4	0.658	0.864	53.3	86.0	0.531	0.780	56.0	55.6	16.9	4.9
	TFR-Net	36.7	84.5	0.801	0.773	50.9	74.7	0.583	0.744	44.4	39.1	61.7	60.7	44.2	86.9	0.676	0.868	46.6	85.3	0.631	0.789	28.8	20.3	47.0	36.9
	GCNet	42.6	85.0	0.740	0.795	51.0	84.8	0.562	0.750	58.7	58.3	59.9	57.9	-	-	-	-	-	-	-	-	-	-	-	-
	IMDer	45.0	84.4	0.717	0.794	51.8	85.0	0.552	0.756	60.4	60.1	59.9	58.2	-	-	-	-	-	-	-	-	-	-	-	-
	MPLMM	37.2	83.9	0.840	0.784	49.8	84.6	0.567	0.761	60.3	59.7	56.5	56.8	-	-	-	-	-	-	-	-	-	-	-	-
	MoMKE	43.7	80.3	0.798	0.756	52.5	85.1	0.544	0.768	60.0	59.9	56.9	56.6	-	-	-	-	-	-	-	-	-	-	-	-
	FuseMoE	43.2	83.2	0.823	0.781	51.0	84.1	0.563	0.737	61.8	61.3	47.2	49.6	-	-	-	-	-	-	-	-	-	-	-	-
	GMD	42.3	86.0	0.770	0.800	53.1	85.1	0.538	0.765	39.5	33.1	57.6	56.5	47.8	87.4	0.614	0.862	54.7	87.4	0.508	0.799	32.9	23.7	59.6	55.1
	EUAR	46.2	85.5	0.723	0.792	52.3	84.4	0.552	0.752	55.7	55.6	58.7	57.2	49.9	89.6	0.572	0.883	52.4	82.6	0.539	0.797	37.7	29.8	59.2	58.8
	LNLN	43.9	83.5	0.761	0.760	52.5	84.3	0.547	0.765	62.8	62.4	58.3	57.1	48.0	88.4	0.601	0.877	54.1	87.2	0.510	0.807	61.4	61.3	54.8	56.3
	P-RMF	43.0	83.3	0.749	0.788	51.3	84.7	0.555	0.766	61.5	60.6	55.8	55.4	-	-	-	-	-	-	-	-	-	-	-	-
$\{a, v\}$	BBFN^r	16.0	42.3	1.385	-0.005	34.0	52.6	0.896	0.452	32.6	21.1	15.8	17.3	-	-	-	-	-	-	-	-	-	-	-	-
	MTMD^r	15.5	42.3	1.415	0.038	41.4	57.5	0.811	0.251	34.7	29.7	48.1	31.3	-	-	-	-	-	-	-	-	-	-	-	-
	ALMT^r	15.6	42.3	1.390	0.054	19.1	48.5	1.379	0.362	30.3	21.1	44.0	34.0	16.5	35.3	1.447	0.120	32.8	39.2	0.922	0.431	32.1	23.9	33.6	31.6
	ITHP^r	15.4	30.4	1.473	-0.002	41.4	51.9	0.843	0.000	17.7	17.8	10.9	2.2	18.4	33.6	1.456	0.128	33.4	41.2	0.897	0.428	31.8	24.5	34.2	32.2
	M3Net^r	-	-	-	-	-	-	-	-	33.7	27.1	48.8	32.5	-	-	-	-	-	-	-	-	-	-	-	-
	CCA	17.4	48.1	1.420	0.117	41.6	63.6	0.806	0.278	39.6	35.1	48.0	36.1	18.7	46.9	1.445	0.102	41.4	68.4	0.798	0.322	40.9	38.1	48.1	39.2
	DCCA	17.5	53.6	1.445	0.093	41.2	73.6	0.790	0.408	35.4	33.3	46.1	33.7	17.4	51.2	1.374	0.172	41.1	71.2	0.792	0.359	38.6	34.5	48.1	31.3
	DCCAE	19.4	55.2	1.569	0.046	37.9	72.7	0.796	0.404	35.5	33.2	46.5	35.7	16.8	52.1	1.368	0.180	41.4	56.9	0.832	0.290	31.4	25.6	48.6	35.1
	CPM-Net	16.8	56.2	1.367	0.330	35.2	72.0	1.395	0.041	44.5	42.4	25.6	28.9	15.5	30.8	1.430	0.095	41.4	48.8	0.834	0.185	28.8	26.3	17.0	14.5
	MM-Align	16.3	56.6	1.358	0.363	38.8	62.6	0.820	0.261	23.7	9.2	2.3	0.2	-	-	-	-	-	-	-	-	-	-	-	-
	NIAT	22.7	42.3	2.056	-0.010	20.9	36.2	1.314	0.129	23.5	8.9	16.9	4.9	17.1	34.4	1.940	0.102	24.1	56.3	1.361	0.070	10.5	2.0	48.4	31.6
	MissModal	15.5	25.1	1.466	0.043	43.4	71.0	0.771	0.405	29.6	22.0	41.3	30.2	-	-	-	-	-	-	-	-	-	-	-	-
	CRA	18.5	55.0	1.410	0.066	40.4	73.2	0.773	0.453	44.1	42.0	45.3	40.1	13.9	30.4	1.679	-0.051	40.9	72.5	0.766	0.436	48.0	46.0	48.5	35.3
	MVAE	17.8	38.7	1.509	0.067	41.0	49.4	0.873	0.182	33.1	33.9	45.1	34.6	-	-	-	-	-	-	-	-	-	-	-	-
	MMVAE	17.3	49.6	1.416	0.103	43.0	50.5	0.813	0.367	30.4	24.8	47.6	32.4	-	-	-	-	-	-	-	-	-	-	-	-
	MCTN	15.5	42.2	1.431	0.018	41.4	62.9	0.842	0.102	22.7	11.5	48.8	31.6	15.5	25.1	1.496	0.016	41.4	48.5	0.841	0.046	23.7	9.1	12.1	2.6
	MMIN	20.3	52.3	1.427	0.081	39.8	71.0	0.774	0.431	48.2	47.4	48.3	31.5	16.5	40.5	1.431	0.149	38.6	72.4	0.784	0.434	44.4	43.8	32.3	32.7
	TFR-Net	18.2	58.0	1.387	0.121	39.2	67.6	0.861	0.414	41.7	40.9	22.1	24.9	17.1	29.5	1.521	0.001	25.1	34.0	1.037	0.429	23.7	9.1	46.1	33.2
	GCNet	17.9	57.2	1.363	0.385	41.2	72.2	0.805	0.393	49.0	48.1	44.7	40.9	-	-	-	-	-	-	-	-	-	-	-	-
	IMDer	18.4	55.1	1.419	0.386	42.0	69.1	0.788	0.376	50.0	49.9	46.7	43.0	-	-	-	-	-	-	-	-	-	-	-	-
	MPLMM	18.2	52.1	1.469	0.137	40.3	72.5	0.770	0.454	37.9	36.2	16.4	16.2	-	-	-	-	-	-	-	-	-	-	-	-
	MoMKE	19.8	54.7	1.459	0.099	42.1	72.9	0.775	0.425	48.4	47.7	13.4	11.6	-	-	-	-	-	-	-	-	-	-	-	-
	FuseMoE	19.0	50.5	1.402	0.262	41.6	72.8	0.769	0.456	45.1	43.8	43.4	32.2	-	-	-	-	-	-	-	-	-	-	-	-
	GMD	17.6	49.1	1.414	0.098	14.8	48.5	1.439	0.427	22.0	14.2	46.1	32.6	21.1	42.3	1.444	0.026	12.6	20.1	1.528	0.380	24.8	11.4	35.1	32.9
	EUAR	15.5	25.7	1.437	0.131	39.9	70.9	0.802	0.427	31.2	26.8	35.4	33.3	20.6	49.8	1.503	0.075	42.4	70.2	0.785	0.385	39.9	39.7	16.9	4.9
	LNLN	15.6	48.7	1.441	0.076	40.8	71.7	0.775	0.441	51.5	50.4	35.5	31.8	15.5	26.1	1.450	0.043	39.9	70.2	0.785	0.383	41.8	41.5	12.1	4.6
	P-RMF	17.4	44.3	1.421	0.063	40.9	73.8	0.766	0.459	40.6	40.6	12.5	7.6	-	-	-	-	-	-	-	-	-	-	-	-

Multimodal Emotion Recognition (MER): *IEMOCAP* (Busso et al., 2008) includes 12 hours of two-way dialogues performed by 10 actors, annotated into six emotions, which are happy, sad, neutral, angry, excited, and frustrated. *MELD* (Poria et al., 2019) comprises approximately 13k utterances from 1,433 multi-party conversations in the TV series Friends, categorized into seven emotion classes including neutral, surprise, fear, sadness, joy, disgust, and angry.

5.2 Evaluation Metrics

For MSA (Poria et al., 2017), the evaluation metrics include seven-class accuracy/non-zero F1-score (*Acc7/F1*) and mean absolute error/Pearson correlation (*MAE/Corr*). *Acc7* measures the average accuracy of sentiment predictions within the label range $[-3, -2.5, -1.5, -0.5, +0.5, +1.5, +2.5, +3]$ loosing the constraints of neutral labels and high intensity, while *F1* evaluates binary positive-negative polarity classification without neutral sentiment. *MAE* quantifies the mean absolute deviation between predictions and ground truth labels, and *Corr* assesses the variation degree of predictions relative to the ground truth.

For MER (A.V. et al., 2024), binary accuracy (*Acc*) and weighted F1-score (*wF1*) are utilized to evaluate the classification performance for overall classes, accounting for the relative quantity of each emotion class to considering label imbalance.

5.3 Comparison Models

We categorize the comparison models as follows:

- **Data augmentation based Models:** The state-of-the-art MAC models under robust training

Table 2: Performance comparison on random missing protocol of the input modalities with missing rates MR set as 0.7 in dataset-level evaluation and MP set as 1.0 in instance-level evaluation.

Language Model		BERT-base / sBERT-base												DeBERTaV3-large									
Random	Models	CMU-MOSI				CMU-MOSEI				IEMOCAP		MELD		CMU-MOSI				CMU-MOSEI				IEMOCAP		MELD	
		Acc7↑	F1↑	MAE↓	Corr↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑	Acc↑	wF1↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑	Acc↑	wF1↑
MR=0.7	BBFN$^{\tau}$	25.3	61.3	1.189	0.430	41.4	56.9	0.794	0.492	36.3	33.1	28.2	32.4	-	-	-	-	-	-	-	-	-	-	-	-
	MTMD$^{\tau}$	26.3	62.2	1.191	0.459	45.2	67.5	0.735	0.447	38.4	37.7	51.2	43.7	-	-	-	-	-	-	-	-	-	-	-	-
	ALMT$^{\tau}$	25.9	61.9	1.169	0.462	32.5	65.1	1.098	0.309	35.5	32.0	50.7	44.2	27.4	55.7	1.176	0.499	34.4	49.2	0.891	0.469	34.7	29.7	48.2	43.5
	ITHP$^{\tau}$	24.9	53.4	1.229	0.441	45.0	64.8	0.748	0.425	30.3	30.5	25.9	27.5	26.3	54.6	1.210	0.452	33.9	48.9	0.856	0.503	25.4	24.8	43.9	43.3
	M3Net$^{\tau}$	-	-	-	-	-	-	-	-	36.5	35.1	50.3	45.4	-	-	-	-	-	-	-	-	-	-	-	-
	CCA	19.9	59.7	1.298	0.326	42.7	68.3	0.775	0.409	41.2	39.8	48.5	37.7	18.9	55.5	1.340	0.269	42.1	69.5	0.785	0.377	38.8	34.8	48.4	37.0
	DCCA	19.6	58.6	1.343	0.214	41.3	61.9	0.806	0.360	36.2	35.0	47.2	32.5	17.0	45.4	1.478	0.070	40.5	63.8	0.828	0.232	28.3	24.4	48.1	31.3
	DCCAE	18.6	55.5	1.539	0.209	38.8	55.9	0.833	0.277	35.7	32.7	47.8	33.2	15.9	36.5	1.439	0.123	41.1	55.1	0.829	0.224	26.3	23.1	48.2	32.1
	CPM-Net	17.1	63.8	1.361	0.291	34.2	73.9	1.692	0.104	52.2	52.5	41.6	34.1	15.5	26.1	1.440	0.087	41.2	49.8	0.847	0.075	27.0	20.4	32.6	31.4
	MM-Align	24.4	60.0	1.274	0.398	37.4	68.8	0.796	0.420	32.5	32.3	20.2	23.8	-	-	-	-	-	-	-	-	-	-	-	-
	NIAT	27.8	59.8	1.651	0.242	23.8	50.6	1.363	0.252	36.7	34.0	31.2	30.7	28.0	55.4	1.299	0.435	43.2	46.8	0.856	0.306	27.9	30.5	51.8	44.9
	MissModal	25.8	52.8	1.206	0.452	45.8	70.6	0.714	0.494	37.1	36.2	32.1	37.0	-	-	-	-	-	-	-	-	-	-	-	-
	CRA	23.6	57.3	1.311	0.401	43.0	76.5	0.713	0.548	40.1	40.0	48.7	42.7	27.3	61.0	1.247	0.372	44.7	76.5	0.7	0.569	45.0	42.8	52.5	44.9
	MVAE	25.4	57.4	1.253	0.410	43.8	59.4	0.791	0.424	35.9	35.9	46.1	41.6	-	-	-	-	-	-	-	-	-	-	-	-
	MMVAE	24.2	59.7	1.218	0.452	44.9	64.7	0.744	0.454	35.5	33.3	50.9	43.2	-	-	-	-	-	-	-	-	-	-	-	-
	MCTN	23.5	56.1	1.223	0.432	43.7	64.9	0.760	0.413	28.8	30.8	51.2	41.2	26.8	54.7	1.179	0.480	45.6	66.2	0.728	0.466	31.1	31.6	21.1	22.0
	MMIN	25.6	59.0	1.212	0.440	43.8	75.3	0.707	0.545	45.9	45.8	51.7	41.9	26.5	60.3	1.179	0.470	43.7	76.0	0.704	0.558	40.8	40.9	23.0	20.3
	TFR-Net	22.8	64.1	1.207	0.400	43.4	62.3	0.782	0.406	37.9	34.6	34.8	39.6	25.2	61.5	1.234	0.367	37.2	52.3	0.877	0.418	25.2	14.4	47.2	34.9
	GCNet	26.3	63.9	1.166	0.494	43.3	74.5	0.733	0.542	47.3	47.2	39.9	41.8	-	-	-	-	-	-	-	-	-	-	-	-
	IMDer	25.0	58.1	1.226	0.426	44.8	69.1	0.732	0.470	47.7	48.2	46.3	43.0	-	-	-	-	-	-	-	-	-	-	-	-
	MPLMM	24.7	64.5	1.251	0.433	42.7	75.0	0.712	0.554	39.3	39.4	27.9	32.3	-	-	-	-	-	-	-	-	-	-	-	-
	MoMKE	26.0	62.4	1.212	0.411	45.8	75.4	0.700	0.540	47.1	46.9	42.8	42.0	-	-	-	-	-	-	-	-	-	-	-	-
	FuseMoE	30.0	55.6	1.436	0.425	45.4	74.7	0.706	0.533	44.8	44.4	44.7	40.7	-	-	-	-	-	-	-	-	-	-	-	-
	GMD	24.7	58.9	1.203	0.468	29.5	65.9	1.123	0.295	24.6	19.7	51.1	43.4	29.9	61.6	1.177	0.433	26.3	50.5	1.203	0.305	26.9	15.8	46.1	41.0
	EUAR	25.6	51.7	1.207	0.446	45.3	72.2	0.717	0.503	37.9	36.5	46.6	42.7	27.9	60.3	1.288	0.411	34.1	61.0	0.925	0.354	32.8	29.7	23.5	29.8
	LNLN	26.1	62.7	1.200	0.472	45.0	74.8	0.705	0.536	46.9	45.9	48.5	43.1	25.8	53.2	1.175	0.491	45.2	71.7	0.711	0.513	41.2	41.5	26.5	28.2
	P-RMF	25.3	55.4	1.209	0.450	44.6	75.6	0.703	0.547	41.6	41.4	26.5	28.5	-	-	-	-	-	-	-	-	-	-	-	-
MP=1.0	BBFN$^{\tau}$	30.4	68.8	1.071	0.548	43.9	66.9	0.726	0.574	42.1	41.0	36.6	40.9	-	-	-	-	-	-	-	-	-	-	-	-
	MTMD$^{\tau}$	31.8	69.5	1.071	0.569	47.4	72.9	0.682	0.547	45.1	45.0	52.9	48.4	-	-	-	-	-	-	-	-	-	-	-	-
	ALMT$^{\tau}$	31.7	68.7	1.050	0.563	37.1	70.5	0.959	0.399	42.5	41.5	52.2	48.0	33.8	66.8	1.000	0.628	40.8	61.1	0.771	0.565	42.8	40.4	49.0	47.0
	ITHP$^{\tau}$	30.6	62.2	1.099	0.556	46.9	70.8	0.696	0.532	37.1	37.2	34.0	37.0	33.1	65.2	1.024	0.605	40.2	60.8	0.752	0.581	36.7	35.5	45.9	43.1
	M3Net$^{\tau}$	-	-	-	-	-	-	-	-	45.7	45.6	53.5	49.6	-	-	-	-	-	-	-	-	-	-	-	-
	CCA	22.2	65.6	1.239	0.403	43.5	72.9	0.745	0.490	46.9	46.2	48.8	40.1	20.0	60.9	1.296	0.342	42.7	72.2	0.761	0.439	44.0	41.3	48.7	39.5
	DCCA	21.8	61.2	1.323	0.296	41.4	68.1	0.799	0.371	42.0	41.3	47.0	33.1	17.2	46.8	1.470	0.088	40.5	65.9	0.817	0.280	32.1	29.0	48.1	31.3
	DCCAE	20.3	60.4	1.541	0.241	38.8	64.0	0.817	0.323	41.0	39.6	47.2	33.7	16.6	40.5	1.429	0.156	41.4	58.8	0.815	0.268	28.4	26.4	48.2	32.5
	CPM-Net	16.9	64.1	1.335	0.339	34.0	73.6	1.400	0.066	52.6	53.1	41.6	34.4	15.5	25.7	1.444	0.076	41.0	50.5	1.255	0.018	26.7	19.4	46.2	34.1
	MM-Align	20.2	61.3	1.418	0.401	39.0	72.0	0.789	0.494	35.6	35.4	23.3	26.8	-	-	-	-	-	-	-	-	-	-	-	-
	NIAT	30.9	67.4	1.426	0.347	31.0	60.7	1.161	0.334	42.7	42.2	38.0	39.1	34.0	66.0	1.111	0.554	44.2	55.6	0.813	0.378	36.3	40.0	53.7	49.5
	MissModal	31.8	63.4	1.060	0.569	47.8	76.0	0.662	0.589	44.4	43.7	39.6	43.4	-	-	-	-	-	-	-	-	-	-	-	-
	CRA	26.9	64.7	1.122	0.505	45.3	78.6	0.671	0.613	47.2	46.8	50.7	46.2	31.3	66.5	1.107	0.504	47.5	79.7	0.639	0.640	50.4	48.3	55.3	49.9
	MVAE	31.0	66.1	1.102	0.531	46.1	66.5	0.727	0.541	42.7	43.1	48.9	45.8	-	-	-	-	-	-	-	-	-	-	-	-
	MMVAE	28.3	67.8	1.095	0.567	46.9	71.3	0.691	0.556	42.6	41.1	52.4	47.8	-	-	-	-	-	-	-	-	-	-	-	-
	MCTN	28.1	63.4	1.115	0.535	44.7	70.3	0.719	0.505	36.4	38.7	52.3	44.3	32.3	64.8	1.031	0.594	47.3	72.4	0.678	0.565	38.5	38.2	25.8	25.6
	MMIN	30.6	67.3	1.092	0.537	46.2	77.9	0.660	0.614	48.4	46.8	54.3	46.4	31.0	67.9	1.041	0.598	46.1	79.1	0.659	0.631	47.3	47.4	25.0	24.6
	TFR-Net	28.1	69.4	1.096	0.510	46.0	70.1	0.719	0.517	44.8	44.0	41.4	46.1	29.6	67.3	1.094	0.528	40.1	62.5	0.794	0.512	31.9	26.1	46.0	35.2
	GCNet	27.6	68.7	1.089	0.549	43.8	78.8	0.742	0.546	54.0	54.1	44.1	45.3	-	-	-	-	-	-	-	-	-	-	-	-
	IMDer	29.5	65.8	1.084	0.541	46.6	74.4	0.689	0.568	55.4	54.9	49.7	46.5	-	-	-	-	-	-	-	-	-	-	-	-
	MPLMM	28.3	70.6	1.127	0.560	44.9	77.9	0.672	0.618	46.9	46.9	36.3	40.6	-	-	-	-	-	-	-	-	-	-	-	-
	MoMKE	30.7	67.1	1.118	0.495	47.3	78.6	0.653	0.612	52.8	52.7	42.2	43.7	-	-	-	-	-	-	-	-	-	-	-	-
	FuseMoE	32.4	66.1	1.152	0.576	46.6	77.2	0.669	0.596	50.9	50.5	45.5	43.9	-	-	-	-	-	-	-	-	-	-	-	-
	GMD	29.4	68.0	1.093	0.561	35.7	71.7	0.976	0.387	32.6	30.3	52.6	47.5	35.4	69.8	1.033	0.549	33.7	61.8	1.020	0.408	35.4	31.7	47.7	44.6
	EUAR	31.6	61.7	1.083	0.548	47.6	76.3	0.665	0.594	42.9	42.5	49.5	47.3	34.5	69.1	1.089	0.542	41.8	71.0	0.781	0.512	40.5	39.9	30.8	39.4
	LNLN	29.8	66.9	1.102	0.531	46.9	77.9	0.662	0.612	54.4	54.9	49.2	46.3	31.3	63.8	1.027	0.612	47.2	77.0	0.656	0.612	48.5	48.9	33.4	36.6
	P-RMF	30.4	65.2	1.075	0.563	46.3	78.4	0.662	0.617	48.6	48.5	34.3	37.2	-	-	-	-	-	-	-	-	-	-	-	-

with random masking data (Hazarika et al., 2022), denoted as $^{\tau}$, including BBFN (Han et al., 2021a), ALMT (Zhang et al., 2023), MTMD (Lin and Hu, 2023b), ITHP (Xiao et al., 2024) for MSA and COGMEN (Joshi et al., 2022), CORECT (Nguyen et al., 2023), M3Net (Chen et al., 2023) for MER;

- **Alignment-based Models:** CCA (Hotelling, 1992), DCCA (Andrew et al., 2013), DCCAE (Wang et al., 2015), MM-Align (Han et al., 2022), CPM-Net (Zhang et al., 2022), NIAT (Yuan et al., 2023), MissModal (Lin and Hu, 2023a);

- **Generation-based Models:** CRA(Tran et al., 2017), MVAE (Wu and Goodman, 2018), MMVAE (Shi et al., 2019), MCTN (Pham et al., 2019), MMIN (Zhao et al., 2021), TFR-Net (Yuan et al., 2021), GCNet (Lian et al., 2023), IMDer (Wang et al., 2023);

- **Adaption-based Models:** MPLMM (Guo et al., 2024), MoMKE (Xu et al., 2024), FuseMoE (Han et al., 2024), GMD (Wang et al., 2024), EUAR (Gao et al., 2024), LNLN (Zhang et al., 2024b), P-RMF (Zhu et al., 2025).

Note that except for CPM-Net (Zhang et al., 2022) and GCNet (Lian et al., 2023) demanding of missing prior in training, all experiments are conducted strictly satisfying the two main principle presented in Section 3. Since MAC highly rely on the performance of text modality due to the rich semantics (Gkoumas et al., 2021), we reproduce all above models with acoustic and visual features extracted by ImageBind (Girdhar et al., 2023),

and textual feature extracted by three widely-used language models with various parameters, including BERT-base (110M) (Devlin et al., 2019) for MSA, sBERT-base (83M) (Reimers and Gurevych, 2019) for MER and DeBERTaV3-large (434M) (He et al., 2021, 2023) for both subtasks. Note that DeBERTaV3-large results are only reported for representative methods that could be feasibly adapted to this stronger encoder. We conduct evaluation on 10 random seeds for all models to denote sufficient missing circumstances and report the average performance as the stable results.

5.4 Performance on Fixed Missing Protocol:

We report experiment results of each model on fixed missing protocol with only textual modality $u \in \{l\}$ and both acoustic and visual modalities $u \in \{a, v\}$ to reveal the contribution of each modality and denote the model behavior on learning unimodal features, as shown in Table 1.

RQ1: What role do different modalities play? Textual modality dominates performance significantly while acoustic/vision modalities remain inferior for all models. When only language is available, models achieve substantially higher performance compared to audio and video only conditions, demonstrating that textual information carries the most discriminative affective signals due to its richer semantic information.

RQ2: How does the scale of language model affect the modality imbalance? Larger language models may amplify modality imbalance rather than mitigate it. DeBERTaV3 shows even greater performance disparity between language-present and language-absent conditions compared to BERT/sBERT. This observation suggests that more powerful language models increase reliance on textual features while contributing less on other modalities, thereby exacerbating the modality imbalance issue in the incomplete scenarios.

RQ3: What type of methods perform more balanced when limited modalities are presented? Alignment-based methods demonstrate superior robustness when limited modalities are presented including NIAT and MissModal. Besides, CPM-Net maintain relatively balanced performance across different fixed missing patterns. While adaption-based methods perform least effectively in language-absent scenarios with severe performance drops when text is missing. This pattern occurs because alignment-based methods explicitly learn to transfer cross-modal synergy information among different modalities, benefiting both MSA and MER subtasks regardless of which modality is available, while adaptation-based methods with limited parameters tend to overfit on the dominant textual modality, achieving strong performance when language is present but degrading significantly when it is absent.

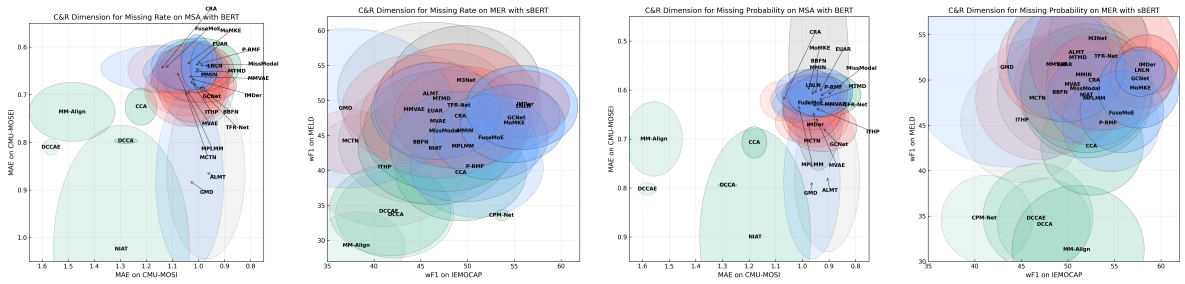
5.5 Performance on Random Missing Protocol:

We report experiment results on random missing protocol with the most severe missing scenarios MR=0.7 and MP=1.0, as shown in Table 1.

RQ1: How does the evaluation level affect the performance? In the most severe missing scenario, instance-level evaluation (MP=1.0) shows consistently better performance than dataset-level evaluation (MR=0.7) across all methods. This discrepancy arises because MR=0.7 simulates missing issue by removing the most modalities in the whole dataset, which causes huger performance degradation across all models with least information provided in each sample. Conversely, instance-level evaluation better matches real-world scenarios where input queries are presented instance by instance with dynamic missing circumstance, making it more practical and less performance drop with MP=1.0 as the most severe scenario.

RQ2: How does larger language model impact the results with random missing scenarios? Larger language models do not consistently improve performance in incomplete multimodal learning due to greater modality imbalance. Although textual modality contributes significantly in MAC task, the drawback of imbalanced contribution among different modalities may largely cause fluctuations when there are incomplete modality inputs, as stronger language encoders underfitting information of inferior modalities and providing less learning on the cross-modal synergy.

RQ3: What type of methods perform better in random missing scenarios? Adaptation-based and generation-based methods contribute most effectively in random missing scenarios benefiting from either adaptive mechanisms dynamically adjusting to varying missing patterns or directly reconstruct missing information as imputation for multimodal fusion. While alignment-based methods can partially enhance incomplete robustness, they struggle with suboptimal performance in the severest missing scenarios due to their reliability on the static fusion representation.



(a) with MR on MSA (b) with MR on MER (c) with MP on MSA (d) with MP on MER

Figure 4: Evaluation on $C\&R$ Dimension with BERT/sBERT on MSA and MER subtasks at both dataset-level (MR) and instance-level (MP) random missing protocols.

5.6 Performance with $C\&R$ Dimension:

The overall performance of each model across both dataset-level and instance-level random missing protocols is presented in Figure 4 with diverse colors denoting four types of models. We observe that adaptation-based methods and generation-based methods consistently cluster in the upper-right region with smaller circle sizes and concentrated positions across all scenarios, indicating they achieve both strong average performance and minimal fluctuation across diverse missing scenarios. While data augmentation based methods and alignment-based methods are dispersed across wider regions with larger axes ranges, which reflects their unstable performance in addressing dynamic missing scenario in the real world.

Besides, models exhibit distinct behavioral patterns between MSA and MER tasks. In sentiment regression of MSA, the performance relies more on the types of model with wider spread, while in emotion classification of MER, models cluster more tightly with generally higher and more similar fluctuation degree. MER demonstrates better overall robustness across diverse model types, with even alignment methods showing reduced fluctuation compared to MSA. This indicates that classification tasks are inherently more resilient to missing modalities than regression tasks where precise sentiment prediction requires more complete multimodal information. The tighter clustering in MER indicates that categorical multimodal emotion features remain more identifiable even with incomplete inputs, whereas continuous sentiment scoring in MSA suffers more dramatically from missing modality issue.

Additionally, models show different stability patterns across MSA datasets, most superior methods exhibit relatively stable performance on MOSEI (smaller vertical axis) but larger fluctuation on MOSI (larger horizontal axis), as evidenced by the ellipses being more horizontally elongated. This suggests that incomplete multimodal learning can benefit from larger-scale dataset with more diverse missing samples to promote generalization ability. While for MER, IEMOCAP's more balanced emotion distribution leads to higher performance fluctuation (larger vertical axis) since models cannot heavily focus on specific dominant emotion classes to improve performance, requiring more robust multimodal understanding across all categories. Conversely, MELD's imbalanced distribution (smaller horizontal axis) allows models to achieve more stable performance by focusing on frequent emotion classes, resulting in reduced fluctuation but potentially less generalization. Thus, more empirical experiments and in-depth analyses are required in the future to address the missing modality issue in multimodal affective computing.

6 Conclusion

In this paper, we present MissMAC-Bench, a comprehensive benchmark for evaluating the missing modality issue in multimodal affective computing. By providing unified evaluation standards to simulate the real world scenarios, we integrate fixed and random missing protocols at dataset and instance levels, which evaluates both competence and resilience of MAC models in complete and incomplete scenarios. Extensive experiments are conducted on 4 types of models across 2 sub-tasks, 3 widely-used language models, and 4 targeted MAC datasets. Our work enables systematic evaluation of inter-modal robustness and cross-modal synergy, offering a solid foundation for developing more generalizable and robust MAC models.

Limitations

Despite its comprehensive design, MissMAC-Bench has several limitations as the future improvement direction. First, while it focuses on modality-level missing scenarios that align with many real-world settings, future extensions could incorporate finer-grained missing patterns such as token-level missing, correlated missing patterns, or noisy input to enable more detailed analysis of intra- and inter-modal dynamics. Second, current benchmark centers on MSA tasks, while expanding the proposed standard to broader multimodal applications could enhance its generalizability. Third, although the benchmark has evaluated several representative encoder-based models in MSA, future work should include more advanced architectures, particularly decoder-based multimodal large language models, to better understand their robustness under missing modality conditions.

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1247–III–1255. JMLR.org.

Geetha A.V., Mala T., Priyanka D., and Uma E. 2024. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, 105:102218.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

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

Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. 2023. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10761–10770.

Zhangtao Cheng, Jian Lang, Ting Zhong, and Fan Zhou. 2025. Seeing the unseen in micro-video popularity prediction: Self-correlation retrieval for missing modality generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, page 142–152, New York, NY, USA. Association for Computing Machinery.

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.

Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Siliang Tang, Kaihang Pan, Yaobo Ye, and 13 others. 2025. On path to multimodal generalist: General-level and general-bench. In *Forty-second International Conference on Machine Learning*.

Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. 2024. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 9650–9659, New York, NY, USA. Association for Computing Machinery.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190.

Dimitris Gkoumas, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. 2021. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197.

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

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021a. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 6–15.

Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. 2022. MM-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10498–10511, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Han, Hui Chen, and Soujanya Poria. 2021b. 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.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. 2024. Fusemoe: mixture-of-experts transformers for fleximodal fusion. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. 2022. Analyzing modality robustness in multimodal sentiment analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–696, Seattle, United States. 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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Harold Hotelling. 1992. *Relations Between Two Sets of Variates*, pages 162–190. Springer New York, New York, NY.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. 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 7042–7052, Online. Association for Computational Linguistics.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States. Association for Computational Linguistics.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14943–14952.

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

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023a. Quantifying & modeling multimodal interactions: an information decomposition framework. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Paul Pu Liang, Yun Cheng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2023b. Multimodal fusion interactions: A study of human and automatic quantification. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 425–435, New York, NY, USA. Association for Computing Machinery.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.* Just Accepted.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*.

Ronghao Lin and Haifeng Hu. 2023a. MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702.

Ronghao Lin and Haifeng Hu. 2023b. Multi-task momentum distillation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–18.

Ronghao Lin and Haifeng Hu. 2024. Adapt and explore: Multimodal mixup for representation learning. *Information Fusion*, 105:102216.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18156–18165.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2302–2310.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: an attentive rnn for emotion detection in conversations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Cam Van Thi Nguyen, Tuan Mai, Son The, Dang Kieu, and Duc-Trong Le. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167, Singapore. Association for Computational Linguistics.

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

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. 2022. Geometric multimodal contrastive representation learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17782–17800. PMLR.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. 2019. *Variational mixture-of-experts autoencoders for multi-modal deep generative models*. Curran Associates Inc., Red Hook, NY, USA.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999.

Ge Teng, Ting Mao, Chen Shen, Xiang Tian, Xuesong Liu, Yaowu Chen, and Jieping Ye. 2024. Urrl-imvc: Unified and robust representation learning for incomplete multi-view clustering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 2888–2899, New York, NY, USA. Association for Computing Machinery.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Miguel Vasco, Hang Yin, Francisco S. Melo, and Ana Paiva. 2022. How to sense the world: Leveraging hierarchy in multimodal perception for robust reinforcement learning agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 1301–1309, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. 2024. Gradient-guided modality decoupling for missing-modality robustness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15483–15491.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1083–1092. JMLR.org.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023. Incomplete multimodality-diffused emotion recognition. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. 2025. On-the-fly modulation for balanced multimodal learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(1):469–485.

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

Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5580–5590, Red Hook, NY, USA. Curran Associates Inc.

Renjie Wu, Hu Wang, and Hsiang-Ting Chen. 2024. A comprehensive survey on deep multimodal learning with missing modality. *arXiv e-prints*, pages arXiv–2409.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2026. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*. Survey Certification.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. 2024. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *The Twelfth International Conference on Learning Representations*.

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

Chengqing Yu, Fei Wang, Chuanguang Yang, Zezhi Shao, Tao Sun, Tangwen Qian, Wei Wei, Zhulin An, and Yongjun Xu. 2025. Merlin: Multi-view representation learning for robust multivariate time series forecasting with unfixed missing rates. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 3633–3644, New York, NY, USA. Association for Computing Machinery.

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.

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*, MM '21, page 4400–4407, New York, NY, USA. Association for Computing Machinery.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2023. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539.

Amir 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.

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*.

Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. 2022. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2402–2415.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. MM-LLMs: Recent advances in MultiModal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand. Association for Computational Linguistics.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024b. 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, Singapore. Association for Computational Linguistics.

Chuang Zhao, Hui Tang, Hongke Zhao, and Xiaomeng Li. 2025. Diffmv: A unified diffusion framework for healthcare predictions with random missing views and view laziness. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 3933–3944, New York, NY, USA. Association for Computing Machinery.

Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep multimodal data fusion. *ACM Comput. Surv.*, 56(9).

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

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 22123–22138, Vienna, Austria. Association for Computational Linguistics.

A Further Discussion with Missing Degree

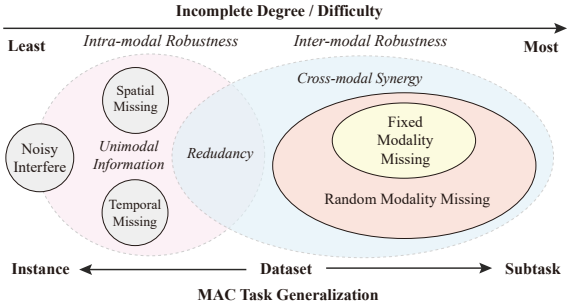


Figure 5: Incomplete multimodal input with diverse missing degrees and task generalization ability.

Based on the missing degree of modality data, we categorize incomplete multimodal input into intra- and inter-modal missing circumstances, as illustrated in Figure 5. The intra-modal missing scenario evaluates the intra-modal robustness to either noisy inputs or spatial and temporal missing scenarios (Yuan et al., 2023), where the unimodal contextual information is partially absent. In contrast, the inter-modal missing scenario simulates cases in which an entire modality is removed from the multimodal input. Intuitively, the inter-modal missing presents greater difficulty, as it requires leveraging cross-modal interactions and compensating for the absence of entire modalities. Besides, inter-modal missing emphasizes the effect of cross-modal synergy, which plays a more complex and crucial role in multimodal fusion (Liang et al., 2023b; Fei et al., 2025) from the perspective of information decomposition (Williams and Beer, 2010; Liang et al., 2023a). Thus in this paper, we focus on scenarios with whole modality missing to reveal model's inter-modal robustness and ability in cross-modal semantic understanding.

Beyond the simulated missingness settings in MissMAC-Bench, we plan to investigate more realistic and temporally structured modality degradation in future work, which further bridge the gap between the proposed missing simulation and the real world scenarios. Specifically, we will introduce modality-specific noise simulators, including misspelling, word-order perturbation, and character insertion for language, muting and background noise for audio, and black-screen, salt-and-pepper noise, Gaussian noise, and blur for visual inputs. We will further consider both fixed noise settings and probability or rate guided instance-level random noise, similar as the evaluation protocols in MissMAC-Bench. This extension will provide a more comprehensive evaluation of multimodal robustness under practical and heterogeneous failure conditions.

B Implementation Details

All experiments are conducted on H800 GPU with CUDA 11.8 and Pytorch 2.0.1. Following (Gkoumas et al., 2021), we perform fifty-times of random grid search for the best hyper-parameters of each model and report the average results over 10 seeds as the final performance.

C Dataset Details

The details about the dataset distribution is introduced as Table 3. We cover MAC task with both MSA and MER subtasks on four datasets. Each subtask contains two size of dataset to reveal partial scalability performance of each model, such as MOSI and MOSEI in MSA. Besides, we consider the impact of class balance in classification for MER subtask, where IEMOCAP dataset has a more balanced distribution on emotion classes while MELD dataset is collected with higher imbalance distribution.

Table 3: Dataset distribution of MissMAC-Bench.

MAC Task	Sentiment Analysis		Emotion Recognition	
Dataset	MOSI	MOSEI	IEMOCAP	MELD
Train	1,284	16,326	5,228	9,765
Valid	229	1,871	519	1,102
Test	686	4,659	1,622	2,524

D More Experiment Results

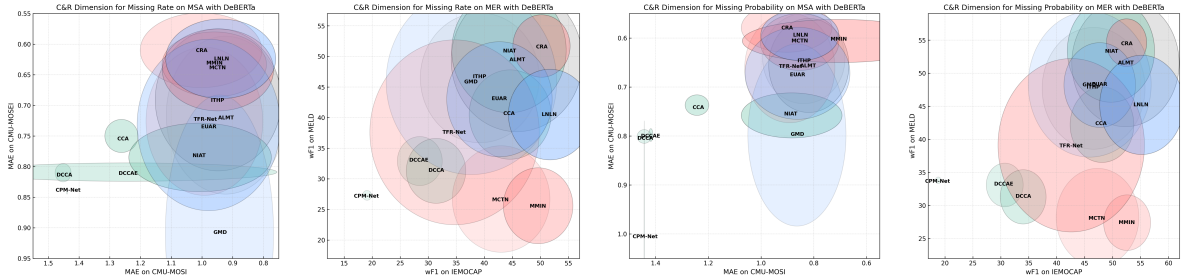
We have released the evaluation code and all experiment results with different missing circumstances specifically in `https://github.com/RH-Lin/MissMAC-Bench`.

D.1 Performance with Complete Modalities

We report experiment results with complete modalities as shown in Table 4. We can observe that performance of diverse models may vary significantly across different datasets. The state-of-the-art methods such as MTMD, ALMT and M3Net mostly reach the highest and balanced performance on

Table 4: Performance comparison with complete input modalities set as $u \in \{l, a, v\}$.

Language Model		BERT-base / sBERT-base												DeBERTaV3-large											
Comp	Models	CMU-MOSI				CMU-MOSEI				IEMOCAP		MELD		CMU-MOSI				CMU-MOSEI				IEMOCAP		MELD	
		Acc7↑	F1↑	MAE↓	Corr↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑	Acc↑	wF1↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc7↑	F1↑	MAE↓	Corr↑	Acc↑	wF1↑	Acc↑	wF1↑
	BBFN^r	45.9	84.5	0.750	0.790	53.9	86.0	0.522	0.788	60.2	60.2	59.6	58.1	-	-	-	-	-	-	-	-	-	-	-	-
	MTMD^r	47.5	85.3	0.705	0.799	53.7	86.0	0.535	0.760	60.1	59.9	57.7	58.0	-	-	-	-	-	-	-	-	-	-	-	-
	ALMT^r	49.0	84.8	0.694	0.806	54.3	84.0	0.519	0.779	63.6	63.5	60.7	59.5	51.5	89.3	0.559	0.884	55.0	86.2	0.494	0.807	67.7	67.7	59.1	58.0
	ITHP^r	44.5	83.6	0.758	0.779	52.0	84.7	0.551	0.756	56.5	56.2	56.6	56.6	50.2	88.5	0.576	0.872	54.2	87.5	0.501	0.810	66.3	66.2	60.2	59.5
	M3Net^r	-	-	-	-	-	-	-	-	59.9	60.3	63.4	61.2	-	-	-	-	-	-	-	-	-	-	-	-
	CCA	27.7	74.9	1.106	0.541	46.1	82.9	0.654	0.666	61.7	61.5	51.2	46.6	24.2	74.9	1.167	0.492	44.0	77.3	0.701	0.565	58.1	57.2	50.5	45.5
	DCCA	25.1	73.6	1.220	0.422	39.3	73.3	0.787	0.425	56.8	55.5	47.7	37.0	16.6	50.4	1.396	0.178	40.8	73.0	0.778	0.406	42.7	40.6	48.1	31.3
	DCCAE	19.7	69.5	1.642	0.357	38.9	73.5	0.782	0.437	57.4	56.4	48.0	36.9	18.1	49.5	1.405	0.186	42.2	68.8	0.786	0.386	37.4	35.6	48.9	34.2
	CPM-Net	16.4	65.5	1.337	0.348	35.9	75.4	0.873	0.375	55.7	56.2	42.3	38.0	15.5	50.7	1.405	0.092	41.4	48.5	0.839	0.132	26.3	18.2	28.2	26.8
	MM-Align	33.5	83.2	0.881	0.717	50.9	82.0	0.570	0.737	49.0	49.4	40.5	36.6	-	-	-	-	-	-	-	-	-	-	-	-
	NIAT	38.8	82.8	0.819	0.746	54.0	85.5	0.529	0.769	62.8	62.7	59.4	60.7	50.9	90.2	0.565	0.883	47.0	75.1	0.679	0.567	62.5	62.6	58.3	58.4
	MissModal	47.1	86.0	0.698	0.801	53.9	85.8	0.533	0.769	63.5	63.7	57.1	57.0	-	-	-	-	-	-	-	-	-	-	-	-
$\{l, a, v\}$	CRA	34.8	83.2	0.916	0.741	51.4	85.5	0.553	0.765	63.4	62.2	57.6	54.8	46.8	87.2	0.615	0.861	55.6	88.3	0.498	0.804	60.8	58.7	62.1	60.6
	MVAE	43.7	84.3	0.745	0.780	52.9	84.4	0.554	0.772	60.9	61.1	54.2	54.5	-	-	-	-	-	-	-	-	-	-	-	-
	MMVAE	39.7	84.2	0.795	0.782	52.2	85.4	0.545	0.770	60.2	59.0	56.6	57.4	-	-	-	-	-	-	-	-	-	-	-	-
	MCTN	43.0	84.6	0.752	0.783	47.9	84.2	0.592	0.721	58.4	57.8	56.3	52.4	48.0	87.9	0.609	0.859	53.4	86.9	0.514	0.798	60.6	59.2	34.2	29.9
	MMIN	43.2	85.0	0.744	0.782	52.9	84.9	0.537	0.769	62.5	62.7	60.6	56.2	45.3	89.5	0.636	0.862	52.5	86.3	0.532	0.790	59.8	59.6	30.2	30.8
	TFR-Net	43.7	86.3	0.774	0.766	55.3	85.8	0.511	0.783	66.8	66.4	59.5	60.8	48.7	88.6	0.608	0.864	57.7	87.6	0.474	0.820	64.6	63.0	45.3	42.9
	GCNet	43.6	85.8	0.732	0.792	52.6	85.9	0.531	0.778	63.0	63.0	60.4	58.5	-	-	-	-	-	-	-	-	-	-	-	-
	IMDer	43.8	85.7	0.724	0.796	52.8	85.1	0.532	0.756	63.4	63.8	61.1	59.7	-	-	-	-	-	-	-	-	-	-	-	-
	MPLMM	37.3	84.2	0.846	0.785	49.8	84.6	0.567	0.761	62.6	62.1	57.1	57.5	-	-	-	-	-	-	-	-	-	-	-	-
	MoMKE	44.5	80.3	0.784	0.758	53.1	85.9	0.528	0.776	64.6	64.6	57.1	56.3	-	-	-	-	-	-	-	-	-	-	-	-
	FuseMoE	45.0	85.2	0.736	0.789	52.0	85.0	0.553	0.747	63.7	63.2	47.1	49.5	-	-	-	-	-	-	-	-	-	-	-	-
	GMD	40.1	85.4	0.790	0.784	54.1	86.5	0.516	0.784	61.3	61.1	57.1	56.3	48.0	88.6	0.612	0.864	55.2	87.9	0.484	0.817	63.0	62.6	53.5	53.4
	EUAR	46.2	85.3	0.725	0.792	54.0	86.1	0.522	0.785	59.9	60.1	58.6	57.5	50.8	89.9	0.566	0.884	54.6	87.6	0.498	0.813	58.3	58.5	59.2	58.8
	LNLN	44.0	84.3	0.762	0.766	52.6	85.1	0.542	0.772	62.9	62.5	58.2	57.1	48.0	88.7	0.601	0.877	54.4	87.5	0.507	0.810	64.2	64.0	54.7	56.2
	P-RMF	43.0	84.0	0.744	0.788	51.8	85.5	0.543	0.775	63.7	63.0	56.0	55.6	-	-	-	-	-	-	-	-	-	-	-	-



(a) With MR on MSA (b) With MR on MER (c) With MP on MSA (d) With MP on MER

Figure 6: Evaluation on $C\&R$ Dimension with DeBERTa on MSA and MER subtasks at both dataset-level (MR) and instance-level (MP) random missing protocols.

both multimodal sentiment analysis and emotion recognition subtasks.

RQ1: Could larger language model achieve better performance? Larger language models demonstrate clear performance advantages in complete multimodal learning, as evidenced by DeBERTaV3 consistently outperforming BERT/sBERT across all metrics. For instance, ITHP achieves 88.5 F1 with DeBERTaV3 compared to 83.6 F1 with BERT on MOSI, representing approximately 5% improvement.

RQ2: How does the data augmentation impact the state-of-the-art complete methods? Compared with the performance reported in the original paper for the state-of-the-art methods such as ALMT or M3Net, leveraging data augmentation with missing modality mostly bring performance degradation on all metrics. We attribute it to the interference caused by the introduction of missing modal data on the fusion space of the original complete modalities.

RQ3: How do different types of models perform when dealing with conflict between complete and incomplete learning? Adaptation-based methods achieve the best balance on both MSA and MER subtasks due to the efficient fine-tuning techniques, while generative-based methods preserve the multimodal fusion weights which can also minimize interference of incomplete input. In contrast, alignment-based methods perform poorly overall in complete multimodal learning due to the large modality gaps, further highlighting the negative impact from incomplete multimodal learning to the complete one.

D.2 $C\&R$ Dimension with DeBERTa

We further provide the $C\&R$ Dimension with DeBERTa as shown in Figure 6. The trends across di-

verse datasets and evaluation levels remain largely consistent with those shown in Figure 4. Nevertheless, we observe that larger language models tend to exhibit greater fluctuations and are highly dependent on the design of different model types, making it more challenging to enhance the robustness of multimodal models as the number of parameters increases. Therefore, research on addressing missing modality issue in the era of large language models still requires further investigation.