Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition

Zirun Guo[1,2]**, Tao Jin**[1*]**, Zhou Zhao**[1,2]
[1] Zhejiang University, [2] Shanghai Artificial Intelligence Laboratory
{gzr,jint_zju,zhaozhou}@zju.edu.cn

Abstract

The development of multimodal models has significantly advanced multimodal sentiment analysis and emotion recognition. However, in real-world applications, the presence of various missing modality cases often leads to a degradation in the model's performance. In this work, we propose a novel multimodal Transformer framework using prompt learning to address the issue of missing modalities. Our method introduces three types of prompts: generative prompts, missing-signal prompts, and missing-type prompts. These prompts enable the generation of missing modality features and facilitate the learning of intra- and inter-modality information. Through prompt learning, we achieve a substantial reduction in the number of trainable parameters. Our proposed method outperforms other methods significantly across all evaluation metrics. Extensive experiments and ablation studies are conducted to demonstrate the effectiveness and robustness of our method, showcasing its ability to effectively handle missing modalities. Codes are available at https://github.com/zrguo/MPLMM.

1 Introduction

Humans perceive the world in a multimodal way, such as sight, sound, touch and language. These multimodal features can provide comprehensive information to help us understand and explore the world. Thus, modeling and mining multimodal data is of great importance and has much potential. Recently, multimodal sentiment analysis (Tsai et al., 2019; Hazarika et al., 2020; Han et al., 2021; Hu et al., 2022) has attracted much attention. However, there are two main challenges in many existing methods: 1) Different from common multimodal tasks which only have two modalities (image and text), multimodal sentiment analysis task often has more modalities (video, audio, text, etc.). There-

fore, in real-world scenarios, missing modality conditions always occur due to equipment failure, data corruption, privacy issues and the like, especially in low-resource domains, which could lead to a degradation in the model's performance. Current multimodal models trained on complete data usually fail when tested on incomplete data (Aguilar et al., 2019; Pham et al., 2019). 2) With the success of large-scale multimodal models (Kim et al., 2021; Li et al., 2021; Radford et al., 2021), lots of researchers tend to finetune these large pre-trained models to downstream tasks. However, this kind of finetuning is infeasible for many researchers because it requires large computational resources. Besides, finetuning such a pre-trained model on small datasets could lead to instability (Mosbach et al., 2021).

Recently, prompt learning (Gao et al., 2021; Heinzerling and Inui, 2021; Khattak et al., 2023; Lee et al., 2023) is proposed, which freezes all the parameters of a pre-trained model while only finetuning several prompts and it has achieved great success (Lester et al., 2021). Motivated by prompt learning, in this paper, we intend to exploit a high-resource dataset that contains relatively more complete modality data for pre-training and then leverage several trainable prompts to transfer the knowledge from high-resource domains to low-resource domains where missing modality cases often occur.

Previous works (Ma et al., 2021; Pham et al., 2019; Zhao et al., 2021) mainly focus on introducing sophisticated architecture to address the issue of missing modalities. These methods do not use pre-trained models and usually require a lot of computational resources. However, our method is based on prompt learning, which only finetunes a few parameters of prompts. Lee et al. (2023) is a recent work which is similar to ours. However, its proposed missing-aware prompts increase exponentially with the number of modalities. In contrast, our proposed prompts increase linearly with the number

---

[*] Corresponding author

of modalities which is more parameter-efficient. Specifically, we propose three types of prompts: generative prompts, missing-signal prompts, and missing-type prompts which can learn the representations of the missing modalities, cross-modal and fine-grained features. These three types of prompts play a combined role in improving the model's performance.

We conduct extensive experiments on four datasets: CMU-MOSEI (Bagher Zadeh et al., 2018), CMU-MOSI (Zadeh et al., 2016), IEMO-CAP (Busso et al., 2008) and CH-SIMS (Yu et al., 2020). The proposed method outperforms the baselines significantly across all metrics on all datasets. We further study the roles of three types of prompts, the effect of missing rate of training data, and the effect of prompt length. We find that: 1) missing-signal prompts are modality-specific while missing-type prompts are modality-shared which represent intra-modality and inter-modality information respectively. 2) with short prompts, our model can achieve very good results which demonstrates our proposed method is parameter-efficient. 3) the missing rate is important for the performance of the model, with 70% being the optimal value.

Our contributions can be summarized as follows:

- We present a novel framework via prompt learning for sentiment analysis and emotion recognition which is not only computationally efficient but also capable of handling missing modalities during both the training and testing stages.
- The number of parameters of our proposed prompts is linearly related to the number of modalities, which significantly reduces computational resources.
- We propose three types of prompts to address the issue of missing modalities. These three types of prompts can generate missing information, and learn intra- and inter-modality information respectively.
- Our proposed method outperforms all the baselines across all metrics significantly. Furthermore, we discover that applying modality dropout with a rate of 70% during training yields the best enhancement in the model's performance.

2   Related Works

Multimodal Sentiment Analysis (MSA) and Emotion Recognition (MER). Multimodal sentiment analysis and emotion recognition refer to the process of analyzing and understanding human sentiment or emotions using multiple modalities of data, such as text, image, audio, and video. The main challenge of such tasks is how to effectively use the information from different modalities to complement each other. Currently, there are two main multimodal fusion strategies: feature-level fusion and decision-level fusion. Feature-level fusion methods (Liang et al., 2018; Wang et al., 2019) combine features from different modalities to create a unified feature representation via concatenation or other methods. For example, Liang et al. (2018) decomposed the fusion problem into multiple stages and fused features step by step to obtain a comprehensive representation. Mai et al. (2019) conducted fusion hierarchically so that both local and global interactions are considered for a comprehensive interpretation of multimodal embeddings. Different from feature-level fusion methods, decision-level fusion methods (Tsai et al., 2019; Hazarika et al., 2020; Han et al., 2021; Hu et al., 2022) process different modalities independently and then incorporate them into the final decision. For instance, Tsai et al. (2019) proposed a directional pairwise cross-modal attention to implement modal alignment and fused the outputs of each modality at the decision level to make predictions. These methods all assume that the data is complete while our proposed method can deal with the situation when there exist missing modalities.

Multimodal Learning with Missing Modalities. The presence of a missing modality poses challenges for multimodal learning because the model needs to effectively handle the absence of information while still making accurate predictions. Ma et al. (2021) proposed the SMIL model which leverages Bayesian meta-learning to address the issue of missing modalities. Some methods (Cai et al., 2018; Du et al., 2018) directly generate missing modalities using the available modalities. Zhao et al. (2021) proposed learning robust joint multimodal representations which can predict the representation of any missing modality given the available modalities. However, these methods always introduced sophisticated architecture to address the issue of missing modalities, which is computationally expensive. In comparison, our approach utilizes three different prompts to handle missing modalities, which is computationally more efficient. In a more recent work (Lee et al., 2023), prompts are used to address missing modalities, but the
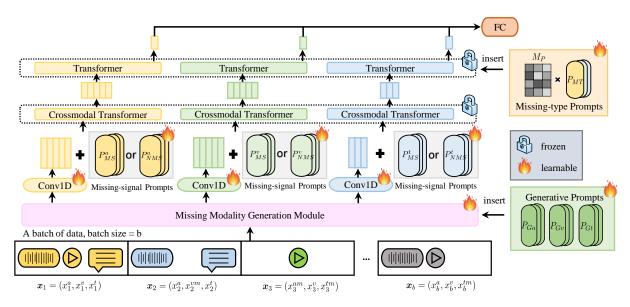
Figure 1: The overall architecture of our proposed method. A batch of data that contains different missing modality cases is fed to the Missing Modality Generation Module (see Section 3.2) to obtain generated features. They are then passed to the pre-trained backbone with missing-signal prompts and missing-type prompts (see Section 3.3).

number of prompts increases exponentially with the number of modalities. In contrast, the number of prompts in our method is linearly related to the number of modalities.

Prompt Learning. Prompt learning, which refers to the process of designing or generating effective prompts to use a pre-trained model for different types of downstream tasks, has been widely used in various NLP tasks (Gao et al., 2021; Heinzerling and Inui, 2021). With the success of prompt learning in NLP tasks (Lester et al., 2021; Li and Liang, 2021; Liu et al., 2022), recent works (Tsimpoukelli et al., 2021; Liang et al., 2022; Khattak et al., 2023) explored to leverage prompts in multimodal learning. Tsimpoukelli et al. (2021) presented a method for transforming large language models into multimodal systems by extending the soft-prompting philosophy of prefix tuning to ordered sets of images and texts. Khattak et al. (2023) proposed a strategy to ensure synergy between vision-language modalities by explicitly conditioning the vision prompts on textual prompts across different Transformer stages. More recently, Lee et al. (2023) proposed missing-aware prompts to address missing modalities which increase the robustness of the model, but it did not recover the missing information from the multimodal input. In comparison, our approach utilizes generative prompts to generate the representation of missing modalities given available modalities which can help further boost the performance of the model.

3 Proposed Method

In this section, we describe our proposed method (Figure 1) via prompt learning to address the issue of missing modalities (introduced in Section 3.1). Specifically, we introduce three kinds of prompts: generative prompts (introduced in Section 3.2), missing-signal prompts, and missing-type prompts (introduced in Section 3.3).

3.1 Overall Architecture

Problem Definition. Given a multimodal dataset $\mathcal{D}$ consisting of $M = 3$ modalities (e.g., audio, video and text), we use $\boldsymbol{x} = (x^a, x^v, x^t)$ to represent a pair of features in $\mathcal{D}$, where x^a, x^v, x^t represent the features of acoustic, visual and textual modalities respectively. To indicate missing modalities, we use x^{am}, x^{vm}, x^{tm} to denote which modalities are absent.

Figure 1 shows the overall architecture of our proposed model. For simplicity and better comparison, we use MulT (Tsai et al., 2019) as the backbone, which introduced the Crossmodal Transformer for modeling unaligned data. In our proposed method, we employ three types of different prompts: generative prompts, missing-signal prompts, and missing-type prompts. The generative prompts assist the available modalities in generating representations for the missing modalities. The missing-signal prompts are designed to inform the model about the absence of a specific modality while the missing-type prompts inform the model
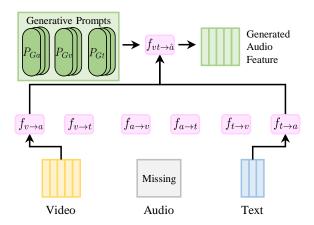
Figure 2: The illustration of Missing Modality Generation Module (MMGM). The figure shows the process of generating the audio feature of an example of $\boldsymbol{x} = (x^{am}, x^v, x^t)$ where the audio modality is missing and the other two are not. It can be described using the Equation 1.
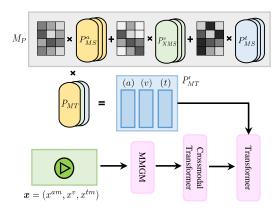

Figure 3: The illustration of attaching missing-type prompts to the Transformer. With the missing-type matrix $\mathbf{M_P}$, we generate missing-type prompts P'_{MT} for different missing modality cases. The figure shows the process of attaching missing-type prompts using an example of $\boldsymbol{x} = (x^{am}, x^v, x^{tm})$ where audio and text modalities are missing.

about the absence of other modalities.

3.2 Missing Modality Generation Module (MMGM)

Many methods address missing modality issues by recovering missing information using available modalities (Cai et al., 2018; Du et al., 2018). However, these methods often utilize complex structures. Based on this observation, we propose the Missing Modality Generation Module (MMGM) which utilizes generative prompts to recover missing information in a much simpler way. We denote generative prompts as $\boldsymbol{P_G} = (P_{Ga}, P_{Gv}, P_{Gt})$ where P_{Ga}, P_{Gv} and P_{Gt} represent the generative prompts for the audio, video and text modalities, respectively. $\boldsymbol{P_G} \in \mathbb{R}^{3 \times d_p \times \ell_p}$ where d_p and ℓ_p represent the dimension and length of the prompts respectively. Figure 2 illustrates the MMGM. Given $\boldsymbol{x} = (x^{am}, x^v, x^t)$, we can generate the representation of the missing modality x^{am} using the available x^v and x^t according to the following equation:

$$\hat{x}^a = f_{vt \to \hat{a}}([P_{Ga}, f_{v \to a}(x^v), f_{t \to a}(x^t)]) \quad (1)$$

where $\hat{x}^a$ denotes the representation generated, $[\ldots]$ represents the concatenation operation, $f(\cdot)$ represents a Conv block which consists of a Conv 1D layer and an activation function and $\to$ represents from one or two modalities to another modality. If there are two missing modalities, such as $\boldsymbol{x} = (x^{am}, x^{vm}, x^t)$, the generation process is as follows:

$$\begin{aligned} \hat{x}^a &= f_{t \to \hat{a}}([P_{Ga}, f_{t \to a}(x^t)]) \\ \hat{x}^v &= f_{t \to \hat{v}}([P_{Gv}, f_{t \to v}(x^t)]) \end{aligned} \quad (2)$$

After applying the MMGM, we can represent the generated features as $\boldsymbol{x} = (\hat{x}^a, \hat{x}^v, x^t)$.

3.3 Missing-signal and Missing-type Prompts

MMGM recovers missing information using available modalities. However, the information generated sometimes might not be accurate and could mislead the model. Therefore, missing-signal prompts are designed to inform the corresponding Transformer whether the information for a particular modality is real or generated. For each modality, there are two missing-signal prompts: P_{MS} to denote a modality is missing and P_{NMS} to denote a modality is not missing. As depicted in Figure 1, after the MMGM and the Conv 1D layer, we obtain features $\boldsymbol{x} = (\hat{x}^a, x^v, x^t)$ where the audio modality is missing originally. We can incorporate the missing-signal prompts as follows:

$$\begin{aligned} \hat{x}^a &:= \hat{x}^a + P^a_{MS} \\ x^v &:= x^v + P^v_{NMS} \\ x^t &:= x^t + P^t_{NMS} \end{aligned} \quad (3)$$

After applying missing-signal prompts, the model knows which modalities are generated and which modalities are real, which can help the model make better use of the recovered information. Notably, missing-signal prompts are modality-specific which means that this kind of prompt only considers a specific modality and does not take into account the correlations between the absence of multiple modalities. To address this limitation, we propose missing-type prompts.

If there are M modalities, there can be a total of $2^M - 1$ different cases of missing modalities. One intuitive approach is to design $2^M - 1$ prompts to handle each situation individually (Lee et al., 2023). However, as the number of modalities increases, this approach becomes computationally expensive. Therefore, we introduce a missing-type projection matrix $\mathbf{M_P}$. We can obtain $\mathbf{M_P}$ of $\boldsymbol{x} = (x^{am}, x^v, x^{tm})$ as follows:

$$\mathbf{M_P} = \mathbf{M_a} \cdot P_{MS}^a + \mathbf{M_v} \cdot P_{NMS}^v + \mathbf{M_t} \cdot P_{MS}^t \quad (4)$$

where $\cdot$ is the matrix multiplication, $\mathbf{M_a}$, $\mathbf{M_v}$, $\mathbf{M_t} \in \mathbb{R}^{d_p \times \ell_p}$ and $\mathbf{M_p} \in \mathbb{R}^{d_p \times d_p}$. Then, we can get the missing-type prompts P'_{MT} as follows:

$$P'_{MT} = P_{MT} \cdot \mathbf{M_P} \quad (5)$$

where P_{MT} represents the original missing-type prompts, P'_{MT} represents the projected missing-type prompts and $P_{MT}, P'_{MT} \in \mathbb{R}^{3 \times \ell_p \times d_p}$. In Figure 3, we illustrate how to attach the missing-type prompts to the Transformer with an example of $\boldsymbol{x} = (x^{am}, x^v, x^{tm})$. For each data pair in $\mathcal{D}$, the corresponding missing-type modality prompt is attached according to the situation of missing modalities.

4 Experiments

4.1 Datasets and Evaluation Metrics

To simulate real-world scenarios, we select CMU-MOSEI (Bagher Zadeh et al., 2018) as the high-resource dataset while CMU-MOSI (Zadeh et al., 2016), IEMOCAP (Busso et al., 2008) and CH-SIMS (Yu et al., 2020) are selected as the low-resource datasets. We pre-train our backbone on CMU-MOSEI and evaluate our proposed method on the four datasets.

CMU-MOSI is a popular dataset for multimodal (audio, text and video) sentiment analysis, comprising 93 English YouTube. Each segment is manually annotated with a sentiment score ranging from strongly negative to strongly positive (-3 to +3).

CMU-MOSEI is an extension of CMU-MOSI. It contains more than 65 hours of annotated video from more than 1000 speakers and 250 topics. Compared with CMU-MOSI, it covers a wider range of topics.

IEMOCAP contains recorded videos from ten actors in five dyadic conversation sessions. It contains approximately 12 hours of data. Following previous works (Wang et al., 2019; Tsai et al., 2019),

four emotions (happiness, anger, sadness and neutral state) are selected for emotion recognition.

CH-SIMS is a Chinese multimodal sentiment analysis dataset. It contains 2,281 video segments annotated with a sentiment score ranging from strongly negative to strongly positive (-1 to 1).

For CMU-MOSI and CMU-MOSEI, we follow previous works and adopt 7-class accuracy (ACC-7), binary accuracy (ACC), F1 score (F1), mean absolute error (MAE) and Pearson correlation (Corr) as evaluation metrics. For IEMOCAP, we implement four binary classification tasks and use the average accuracy (ACC) and F1-weighted score (F1) as evaluation metrics. For CH-SIMS, we use binary accuracy (ACC), F1 score (F1), mean absolute error (MAE) and Pearson correlation (Corr).

4.2 Baselines

We compare our proposed method with the following methods: **Lower Bound (LB)** is trained with different combinations of modalities. Specifically, we train six different models using different combinations of modalities. **Modality Substitution (MS)** substitutes missing modality with a default value or a placeholder. **Modality Dropout (MD)** is a model trained with randomly dropped modalities during the training phase. **MCTN** (Pham et al., 2019) learns robust joint representations by translating between modalities to deal with missing information. **MMIN** (Zhao et al., 2021) learns robust joint multimodal representations, which can predict the representation of any missing modality given available modalities. **MPMM** (Lee et al., 2023) uses missing-aware prompts to instruct the model to address missing modality issues.

4.3 Implementation Details

Raw Feature Extraction. To demonstrate the generalization ability of our method, we implement three kinds of methods to extract features. For CMU-MOSEI and CMU-MOSI, we follow Tsai et al. (2019) to extract features. For IEMOCAP, we follow Zhao et al. (2021) to extract acoustic, visual and textual features. For CH-SIMS, we follow Yu et al. (2020) to extract features.

Model Training Details. We first pre-train our backbone MulT (Tsai et al., 2019) on the CMU-MOSEI dataset. Then, we freeze all the parameters of the backbone and only train several learnable prompts, Conv layers and the output layer (As shown in Figure 1). The length of prompts ℓ_p is set to 16 by default. We use L1 loss function for CMU-

Table 1: Quantitative results under six possible missing modality cases. For example, "$\{a\}$" means audio modality is available while video and text are missing. "Avg." means the average performance of the six possible cases. † denotes results copied from Zhao et al. (2021) where F1 score is not reported. **Bold**: best result. Underline: second best result. We report the average result of five different random seeds.

Dataset	Method	{a} ACC	{a} F1	{v} ACC	{v} F1	{t} ACC	{t} F1	{a,v} ACC	{a,v} F1	{a,t} ACC	{a,t} F1	{v,t} ACC	{v,t} F1	Avg. ACC	Avg. F1
MOSI	LB	48.32	55.81	49.09	55.20	79.27	79.22	50.07	57.12	78.67	79.25	79.86	79.96	64.21	67.76
	MS	49.17	55.34	49.87	56.12	78.06	78.28	51.12	57.01	79.32	79.65	80.32	80.38	64.64	67.80
	MD	48.79	55.74	49.66	55.60	79.36	80.01	52.33	56.84	79.59	79.86	80.51	80.43	65.04	68.08
	MCTN	51.32	56.12	54.27	56.33	79.63	79.78	56.79	57.84	78.96	79.17	80.45	80.65	66.90	68.32
	MMIN	59.16	60.12	61.01	61.98	80.10	80.16	63.79	64.08	80.50	80.33	80.46	80.63	70.84	71.22
	MPMM	57.26	59.35	58.63	59.12	79.81	80.10	60.54	61.33	79.89	79.84	80.74	80.93	69.48	70.11
	Ours	62.71	63.65	63.12	63.74	80.12	80.31	65.02	65.41	80.76	81.09	81.12	81.19	72.14	72.57
IEMOCAP	LB	46.35	46.21	48.07	47.58	56.06	55.28	58.12	57.89	72.18	72.25	65.63	65.28	57.74	57.42
	MS	47.65	47.52	47.68	47.36	59.27	59.22	57.48	56.60	72.30	72.18	66.81	66.93	58.53	58.30
	MD	48.22	48.09	48.26	47.98	61.26	61.28	58.08	57.96	72.40	72.31	67.08	68.22	59.22	59.31
	MCTN	51.62†	-	45.73†	-	63.78†	-	55.84†	-	69.46†	-	68.34†	-	59.19†	-
	MMIN	59.00†	-	51.60†	-	68.02†	-	65.43†	-	75.14†	-	73.61†	-	65.47†	-
	MPMM	58.69	57.66	55.18	55.36	68.39	68.08	63.68	63.47	74.90	74.98	73.80	72.67	65.77	65.37
	Ours	59.77	59.71	57.61	56.98	69.23	69.28	67.26	67.37	75.98	75.44	74.68	74.51	67.42	67.22
CH-SIMS	LB	63.82	75.15	64.08	78.11	76.74	76.90	62.14	73.21	76.84	76.93	77.01	77.13	70.11	76.24
	MS	62.45	74.59	63.58	76.86	77.28	77.84	60.18	71.09	76.01	76.30	77.13	77.20	69.44	75.65
	MD	64.22	77.25	63.87	76.01	77.34	77.48	62.91	72.14	76.77	76.92	77.14	77.31	70.38	76.19
	MCTN	64.39	76.48	64.12	76.34	77.78	77.92	63.47	73.11	76.68	76.71	77.21	77.36	70.61	76.32
	MMIN	65.21	77.09	65.32	77.41	78.91	78.67	64.28	73.36	77.32	77.33	77.40	77.48	71.41	76.89
	MPMM	64.98	76.41	65.40	77.92	78.56	78.65	64.01	73.47	77.11	77.20	77.51	77.47	71.26	76.85
	Ours	65.93	77.10	66.02	78.86	79.75	78.74	65.28	74.02	77.45	77.84	77.97	77.95	72.07	77.42
MOSEI	LB	66.21	68.69	66.45	69.10	77.96	78.32	67.30	69.62	78.13	78.63	77.86	78.16	72.32	73.83
	MS	62.74	67.06	64.16	68.17	77.28	77.76	67.11	69.51	78.34	78.80	78.08	78.32	71.29	73.36
	MD	65.76	68.18	66.57	69.35	77.30	77.94	67.21	69.48	78.74	78.97	78.11	78.71	72.28	73.82
	MCTN	66.19	68.58	66.70	69.01	78.32	78.41	68.10	69.34	79.11	79.14	78.65	78.64	72.85	73.94
	MMIN	67.11	68.67	67.01	69.31	78.67	78.71	68.17	69.74	79.94	79.96	79.32	79.29	73.37	74.39
	MPMM	66.94	68.74	67.21	69.27	78.21	78.30	68.11	69.79	79.41	79.47	79.63	79.71	73.25	74.17
	Ours	67.33	68.71	67.29	69.40	79.12	79.17	68.21	69.91	80.45	80.43	80.11	80.13	73.75	74.68

MOSEI, CMU-MOSI and CH-SIMS datasets and cross-entropy loss for IEMOCAP dataset. In all experiments, we use Adam optimizer with a batch size of 64. We train the model for 30 epochs with a learning rate of 1×10^{-3}. Besides, we randomly discard the modality of the data with a missing rate of $\eta = 70\%$ during training. We fix the random seed to ensure that each model is trained on the same data.

4.4 Main Results

In Table 1, we present quantitative results on four datasets. The baselines LB, MS, MD and MPMM share the backbone of our method, thus reflecting the effectiveness of our proposed method to deal with missing modalities. Comparing the baseline MS and MD, we find that random discarding of data modalities during training improves the generalization ability of the model, thus making the model less sensitive to the data with missing modalities during the test phase.

Analyzing the results presented in the table, we observe that our proposed method outperforms the baselines by a large margin in all datasets under all six missing modality cases. Additionally, our method brings great enhancement when text modality is missing, with 8-13% increase in accuracy compared with the LB baseline. This indicates that the three types of proposed prompts effectively guide the pre-trained model and yield impressive performance improvements.

Besides, it is worth noting that we implement different feature extraction approaches on four datasets. From the results in Table 1, we can see that our model outperforms the baselines on all datasets, which shows that our model can adapt to features extracted by different methods. This indicates that our model learns relative rather than absolute relationships between features, demonstrating its robustness and versatility.

In Figure 4, we further compare the performance of our model with other methods under different modality missing rates during test time. From the figure, we find that our model performs better than all the other methods across all metrics under different modality missing rates, although the model
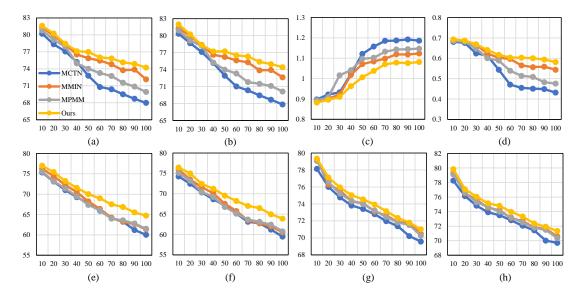
Figure 4: Performance comparison with different modality missing rates during tests. (a): ACC on CMU-MOSI. (b): F1 score on CMU-MOSI (c): MAE on CMU-MOSI. (d): Corr on CMU-MOSI. (e): ACC on IEMOCAP. (f): F1 score on IEMOCAP. (g): ACC on CH-SIMS. (h): F1 score on CH-SIMS.

is trained using the dataset with a modality missing rate of $\eta = 70\%$. This indicates that our proposed method is robust to the missing rate of test set and can deal with severely missing modalities well.

Furthermore, the number of trainable parameters of our method is about 5-10% percent of that of the backbone. The majority of trainable parameters come from the Conv layers in MMGM. The number of trainable parameters of three types of prompts only accounts for 0.5-1% of that of the backbone. Notably, the number of trainable parameters does not increase with the size of the backbone network, which means that even if we use a much larger backbone, the number of trainable parameters remains the same. In all our experiments, we use only one 10 GB GPU (RTX 3080) with a batch size of 64. This demonstrates that our method is parameter-efficient.

4.5 Generalization Ability

To further validate the generalization ability of our method, we conduct experiments using different MSA/MER backbones. Specifically, we conduct experiments using MISA (Hazarika et al., 2020), MMIM (Han et al., 2021) and UniMSE (Hu et al., 2022) and present the results in Table 2. For all three backbones, we insert our generative prompts and module after the feature extractors. For UniMSE, we insert missing-signal and missing-type prompts into its multimodal fusion layers. For MMIM and MISA, we insert missing-signal and missing-type prompts into their modality-specific

Table 2: Performance on different backbones on the CMU-MOSI dataset. "Com" denotes the complete data. "Incom" denotes the incomplete data. † denotes the method attached with prompts. For incomplete data, we report the average accuracy of six different missing conditions.

Backbone	Com	Com†	Incom	Incom†
MISA	80.8	$81.4_{(+0.6)}$	67.9	$73.4_{(+5.5)}$
MMIM	84.1	$84.9_{(+0.8)}$	68.6	$72.3_{(+3.7)}$
UniMSE	86.9	$87.4_{(+0.5)}$	69.8	$75.1_{(+5.3)}$

encoders and fusion layers, respectively. The results in the table demonstrate our method can enhance the ability of various backbones to address missing modality issues. Besides, our prompts can enhance the performance in the complete data situation, indicating that our missing-signal and missing-type prompts can help the model learn intra-modality and inter-modality features.
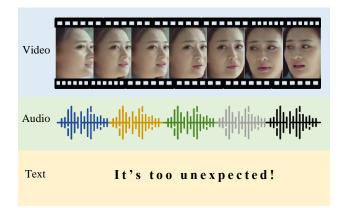
4.6 Ablation Study

We divide our ablation experiments into three parts: contributions of three types of prompts, the effect of modality missing rate during training and the effect of prompt length.

Contributions of three types of prompts. In Table 3, we present quantitative results of the contributions of different prompts. For CMU-MOSI, we can observe that generative prompts give the greatest improvement in ACC and F1, while missing-signal prompts improve the Corr the most and missing-type prompts improve the ACC-7 the most.

Table 3: An ablation study on the benefit of the proposed generative prompts P_G, missing-signal prompts P_{MS} and missing-type prompts P_{MT}. ✓ represents a model with such type of prompts. **Bold**: best results with two kinds of prompts attached. * denotes best results with only one kind of prompt attached. We report the average performance of the six possible missing modality cases.

P_G	P_{MS}	P_{MT}	CMU-MOSI					CH-SIMS				IEMOCAP	
			ACC-7	ACC	F1	MAE	Corr	ACC	F1	MAE	Corr	ACC	F1
			27.41	65.04	68.08	1.121	0.549	70.38	76.19	0.555	0.506	59.22	59.31
✓			28.27	70.26*	71.33*	1.097*	0.574	71.08*	76.57*	0.518*	0.521	64.12*	63.97*
	✓		28.11	67.62	69.40	1.104	0.580*	70.61	76.21	0.531	0.527*	62.27	62.31
		✓	30.17*	66.41	69.21	1.167	0.561	70.49	76.13	0.542	0.515	63.48	63.20
✓	✓		30.63	**71.94**	**72.10**	1.084	0.576	**71.84**	**77.13**	**0.496**	0.536	65.23	65.01
✓		✓	31.27	71.60	71.86	**1.075**	0.583	71.36	76.95	0.498	0.533	**66.11**	**66.04**
	✓	✓	**32.88**	67.56	69.31	1.091	**0.593**	70.92	76.34	0.521	**0.541**	64.92	64.67

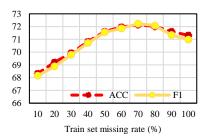|			P_G	P_{MS}	P_{MT}	Prediction
Video						Positive
			✓			Weakly Positive
				✓		Weakly Positive
Audio					✓	Weakly Positive
			✓	✓		Weakly Negative
			✓		✓	Negative
Text	It's too unexpected!			✓	✓	Neutral
			✓	✓	✓	Negative

Figure 5: The effectiveness of three types of prompts on an example of CH-SIMS. The ground truth of the sample is "Negative". We report the results when the visual modality is missing.
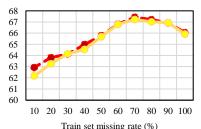
This indicates that generative prompts can help the available modalities generate the missing information which improves the binary accuracy. Besides, missing-type prompts tell the model whether other modalities are missing, thus strengthening the interactions between different modalities and learning cross-modal and fine-grained information which helps improve the ACC-7 a lot.

From the performance of models with different combinations of three types of prompts, we can further demonstrate the different roles of three types of prompts. We can conclude that generative prompts learn good representations of the missing modalities and improve the binary accuracy, missing-signal prompts are modality-specific prompts that tell models whether the corresponding modality is missing and help improve the correlation of the model's predictions with humans, and missing-type prompts are shared prompts with inter-modality information, thus helping models learn cross-modal and fine-grained information that improves ACC-7. Furthermore, the combinations of three types of prompts further enhance the performance of the model on all datasets. This fully

confirms the validity of our proposed method. Besides, we use an example in CH-SIMS to study the effectiveness of three types of prompts and present the results in Figure 5. From the figure, we can observe that the visual modality is key to predicting the correct result. With the prompts attached, the model can predict the accurate result, indicating the effectiveness of our method.

The effect of modality missing rate during training. We study the impact of modality missing rate during training on the performance of the model in Figure 6. From the figure, we find that starting at a low point, both ACC and F1 score steadily improve as the train set modality missing rate increases, before reaching the highest point when the missing rate $\eta = 70\%$. Then both ACC and F1 score decrease as the missing rate increases. This indicates that when the train set missing rate is low, it is difficult for a model to learn very good representations in the MMGM and to learn opportune prompts that can instruct the model well. This is because when the missing rate is low, the model tends to find a shortcut which to some degree prevents the model from learning good representations.
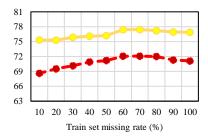
Figure 6: Quantitative results on CMU-MOSI (left), IEMOCAP (middle) and CH-SIMS (right) with different modality missing rates during training. We report the average performance under six different missing cases.
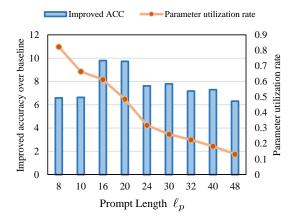

Figure 7: Quantitative results on CMU-MOSI with different prompt lengths ℓ_p. The figure shows the improved accuracy (IACC) over the baseline Modality Dropout and the parameter utilization rate $\xi = \text{IACC}/\ell_p$.

With the missing rate higher, the model has to learn how to generate missing information to make predictions more accurate. However, if the missing rate is higher than 70%, due to the amount of missing information, it is also hard for a model to learn good representations and prompts.

The effect of prompt length. To study the impact of prompt length on our model, we train our model on CMU-MOSI with nine different prompt lengths and present results in Figure 7. In the figure, we show the improved accuracy (IACC) over the baseline "Modality Dropout" of models with different prompt lengths. Intuitively, the longer the prompt length, the better the performance of the model. However, with the results shown in the figure, we find that when the prompt length $\ell_p = 16$, the model performs the best. When the prompts are longer than 20, with the increase of the prompt length, the performance of the model decreases. Therefore, we deduce that it may be because our task is not complex and therefore the increase in parameters may overfit the model. Besides, we introduce parameter utilization rate $\xi = \text{IACC}/\ell_p$ to represent a trade-off between the performance of

models and the number of parameters of prompts. From the figure, we can clearly see that $\ell_p = 16$ is the best choice, where IACC and ξ are both high compared with others. This also indicates that our proposed method can help improve the baseline with only a few parameters.

5 Conclusion

In this paper, we propose a novel multimodal Transformer via prompt learning to tackle the issue of missing modalities. We propose three types of prompts: generative prompts, missing-signal prompts, and missing-type prompts. Generative prompts can help generate missing information. Missing-signal prompts are modality-specific and missing-type prompts are modality-shared, which help the model learn intra-modality and inter-modality relationships respectively. With prompt learning, we can significantly reduce the number of trainable parameters. Extensive experiments and ablation studies demonstrate the effectiveness and robustness of our proposed method.

Limitations

Our missing modality generation module generates missing information through two simple Conv blocks and generative prompts. This module improves the performance of our model significantly. However, we use extracted features but not raw features. Due to the simplicity of our missing generation module, the performance of the model could degrade if we use raw features which are much more complicated and have weaker correlation between modalities than extracted features. We leave this problem to future work.

Acknowledgements

References

Gustavo Aguilar, Viktor Rozgic, Weiran Wang, and Chao Wang. 2019. Multimodal and multi-view models for emotion recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 991–1002, Florence, Italy. Association for Computational Linguistics.

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.

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

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

Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. 2018. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM international conference on Multimedia*. ACM.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. 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.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries.

In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

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*, pages 14943–14952.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161, Brussels, Belgium. Association for Computational Linguistics.

Sheng Liang, Mengjie Zhao, and Hinrich Schuetze. 2022. Modular and parameter-efficient multimodal

fusion with prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2976–2985, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

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.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492, Florence, Italy. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

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.

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

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.