Enhancing Multimodal Model Robustness Under Missing Modalities via Memory-Driven Prompt Learning

Yihan Zhao, **Wei Xi** *, **Xiao Fu** and **Jizhong Zhao**
School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
zhaoyihan@stu.xjtu.edu.cn, xiaofu9804@gmail.com, {xiwei, zjz}@xjtu.edu.cn

Abstract

Existing multimodal models typically assume the availability of all modalities, leading to significant performance degradation when certain modalities are missing. Recent methods have introduced prompt learning to adapt pretrained models to incomplete data, achieving remarkable performance when the missing cases are consistent during training and inference. However, these methods rely heavily on distribution consistency and fail to compensate for missing modalities, limiting their ability to generalize to unseen missing cases. To address this issue, we propose Memory-Driven Prompt Learning, a framework that adaptively compensates for missing modalities through prompt learning. The compensation strategies are achieved by two types of prompts: generative prompts and shared prompts. Generative prompts retrieve semantically similar samples from a predefined prompt memory that stores modality-specific semantic information, while shared prompts leverage available modalities to provide cross-modal compensation. Extensive experiments demonstrate the effectiveness of the proposed model, achieving significant improvements across diverse missing-modality scenarios, with average performance increasing from 34.76% to 40.40% on MM-IMDb, 62.71% to 77.06% on Food101, and 60.40% to 62.77% on Hateful Memes. The code is available at https://github.com/zhao-yh20/MemPrompt.

1 Introduction

The field of multimodal learning has made remarkable progress, primarily driven by the capability of transformer to effectively integrate diverse modalities [Akbari *et al.*, 2021; Nagrani *et al.*, 2021]. These models exploit the complementarity and associations among modalities to achieve excellent performance in tasks such as image captioning, visual question answering, and cross-modal retrieval [Shao *et al.*, 2023; Stefanini *et al.*, 2022; Wang *et al.*, 2016]. Despite these advancements, a critical challenge remains: the performance of multimodal models degrades substantially when one or more modalities are missing. Existing multimodal models typically assume that all modalities are available during both training and inference [Kim *et al.*, 2021]. In real-world applications, incomplete data is common due to sensor failures or privacy concerns. This reliance on complete data limits their robustness and impedes their deployment in practical scenarios.

Previous researches on handling missing modalities primarily focus on multi-stream multimodal architectures, which often reconstruct missing modality features based on available modalities [Ma *et al.*, 2021; Wang *et al.*, 2023; Woo *et al.*, 2023]. These paradigms require retraining the model on incomplete data to enable the model to cope with missing modalities scenarios. Recently, multimodal models based on unify transformers have demonstrated powerful performance. [Ma *et al.*, 2022] first explored robustness against missing modalities in unified multimodal transformer, and introduce multi-task training approach to improve model performance. As the model scale continues to grow, some researcher introduced the parameter efficient finetuning to copy with missing modalities cases for avoiding expensive computation cost. [Lee *et al.*, 2023] designed specific prompts tailored to different missing-modality cases, while [Shi *et al.*, 2024] introduce modality-aware adapter to disease diagnosis for handling missing scenarios.

However, there are still several limitations. First, they fail to leveraging existing samples and available modalities to compensate for the missing modality information. Unlike multi-stream architectures, unified transformer model concatenate all modality features into a unified representation, which makes it challenging to disentangle individual modalities. This constrains the ability to reconstruct missing modality representations using existing ones. Second, current approaches focus on transferring pretrained models to missing modality scenarios, resulting in failing to handle unseen missing scenarios during training. These methods often rely on customized prompts tailored to specific missing scenarios during training. As a result, they rely heavily on pre-defined missing scenarios, making them inflexible when faced with unseen combinations of missing modalities during inference. For instance, the training data is only accompanied with missing text modality, the model performance deteriorates significantly when the visual modality is missing during inference. Therefore, in incomplete data scenarios, it is crucial to fully

*Corresponding author.

utilize the available information to compensate for missing information, thereby enhancing the multimodal model's robustness and its ability to handle diverse missing scenarios.

To address the above limitations, we propose Memory-Driven Prompt Learning (**MemPrompt**), an approach that adaptively compensates for missing modality information to enhance model generalization and robustness. Due to the difficulty of disentangling individual modality representations in multimodal transformer architectures, we design a prompt-based strategy that leverages two complementary information sources: (1) semantically similar samples from the current modality, and (2) shared information from the available modalities. To achieve this, we introduce two types of prompts: generative prompts and shared prompts. Generative prompts are dynamically retrieved from a predefined prompt memory, which stores semantic information of all samples for each modality, enabling the retrieval of semantically similar samples to compensate for missing information. Additionally, shared prompts establish cross-modal associations, allowing existing modalities to provide relevant information for missing modalities. Extensive experiments demonstrate that our method achieves state-of-the-art performance, regardless of whether the missing modalities during training and inference are consistent or not. This is attributed to the effectiveness of our compensation mechanism, which dynamically leverages both intra-modal and cross-modal information.

The main contributions can be summarized as follows:

- We propose Memory-Driven Prompt Learning for multimodal models to handle diverse missing modality scenarios. This approach employs a prompt-based strategy to compensate for missing information.
- The missing information is compensated through two complementary prompts: generative prompts, which retrieve semantically similar features from a predefined prompt memory, and shared prompts, which utilize cross-modal associations to supplement the missing information.
- Extensive experiments on three benchmark datasets demonstrate the superior performance of the proposed method, validating its effectiveness and adaptability, even under inconsistent missing modality scenarios.

2 Related Work

2.1 Multimodal Learning with Missing Modalities

Multimodal learning aims to fuse diverse modalities to leverage richer information, thereby enhancing model performance [Kim *et al.*, 2021; Akbari *et al.*, 2021; Radford *et al.*, 2021; Bao *et al.*, 2022]. Conventional approaches typically assume that all modalities are available during both training and inference. However, in real-world applications, certain modalities may be unavailable, resulting in model performance significantly deteriorates[Ma *et al.*, 2021; Ma *et al.*, 2022; Wang *et al.*, 2023; Shi *et al.*, 2024]. Consequently, many researchers have focused on improving the robustness of multimodal models in scenarios with missing modalities [Zhao *et al.*, 2024b; Reza *et al.*, 2024; Yang *et al.*, 2024; Wu *et al.*, 2024b]. Existing methods attempt to simulate the missing modality data during training so that the model can generalize to the missing modalities scenarios. They primarily focus on reconstructing missing modality features from available ones. For instance, [Woo *et al.*, 2023] proposed ActionMAE, which reconstructs features of missing modalities using the remaining modalities. Similarly, [Wang *et al.*, 2023] introduced Shared-Specific Feature Modeling, which improves robustness by learning both shared and modality-specific features. However, the above approaches require to retrain the model, which undoubtedly increases the computational overhead. Some methods generalize the model to different missing scenarios through parameter-efficient fine-tuning. [Reza *et al.*, 2024] proposed parameter-efficient adaptation techniques to compensate for missing modalities. [Lee *et al.*, 2023] introduced missing-aware prompts to generalize pretrained models to handle missing modality scenarios, while MoRA [Shi *et al.*, 2024] employs distinct modality-aware up-projections within LoRA [Hu *et al.*, 2021] to enable modality-specific adaptations. However, these approach fail to employ existing information to compensate for missing modalities, thus resulting in performance drop under inconsistent missing cased during training and inference.

2.2 Prompt Learning

As the scale of transformer-based models continues to grow, the cost of fully fine-tuning these pretrained models is becoming increasingly prohibitive. Prompt learning has emerged as an efficient paradigm that leverages task-specific prompts to guide pretrained models, enabling them to adapt to downstream tasks with minimal fine-tuning while fully utilizing the knowledge embedded within the pretrained models [Li and Liang, 2021; Jia *et al.*, 2022; Petrov *et al.*, 2024]. This approach has been widely adopted in visual tasks [He *et al.*, 2023a; Han *et al.*, 2023] and multi-modal learning [Roy and Etemad, 2024; Park *et al.*, 2024; Xin *et al.*, 2024b; Khattak *et al.*, 2023a; Zhao *et al.*, 2024a]. Visual prompt learning [Jia *et al.*, 2022; He *et al.*, 2023b; Han *et al.*, 2023; Wang *et al.*, 2024] introduces a small number of trainable parameters at the input layer while keeping the pretrained model frozen. In the multi-task learning domain, works such as [Xin *et al.*, 2024a; Xin *et al.*, 2024b] have employed prompt learning to align text and visual modalities during the fine-tuning process. In multi-modal learning, studies like [Zhou *et al.*, 2022; Khattak *et al.*, 2023a] utilize prompt learning to enhance few-shot learning performance. Furthermore, researchers have addressed the issue of over-fitting in prompt learning [Khattak *et al.*, 2023b; Roy and Etemad, 2024] and explored the integration of parameter-efficient fine-tuning (PEFT) with expert systems to enhance the model's representation capabilities [Li *et al.*, 2024; Wu *et al.*, 2024a]. In this work, we employ a prompt learning strategy to address missing modality scenarios by compensating for the absent information.

3 Method

3.1 Problem Definition

For simplicity and without loss of generality, we consider a multimodal dataset consisting of two modalities, denoted as

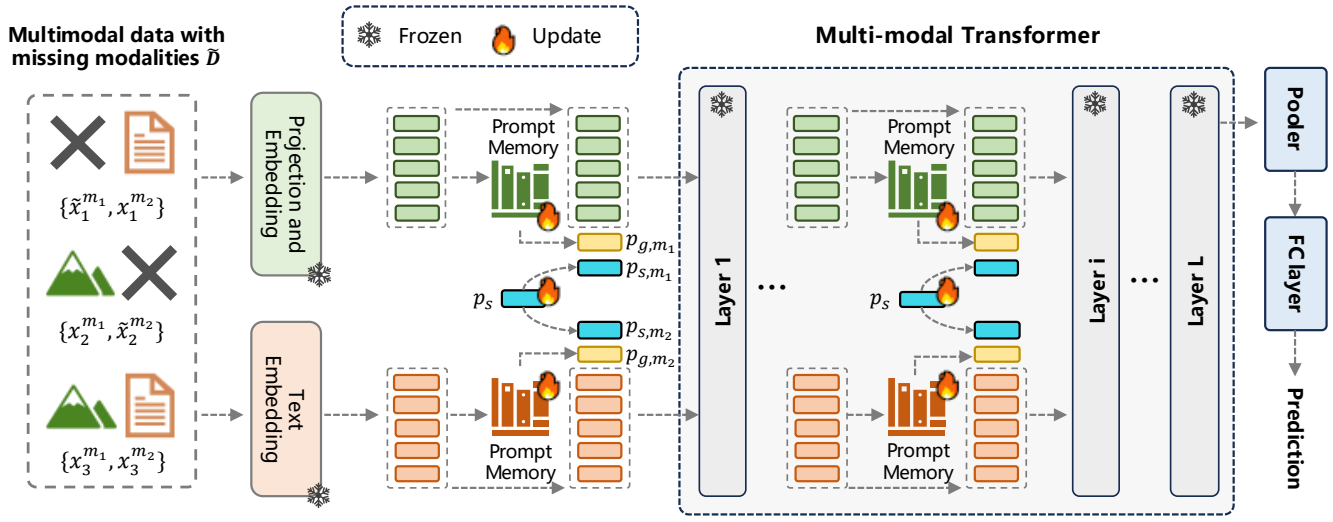


Figure 1: An overview of the proposed framework. The model incorporates two types of prompts: generative prompts and shared prompts. Generative prompts are adaptively retrieved from a modality-specific prompt memory, while shared prompts are mapped to each modality via a mapping function. The prompts are directly added to the corresponding modality features, while the pretrained backbone remains frozen, with only the prompts and task-specific layers updated.

$M = 2$, with modalities m_1 and m_2. The dataset is defined as $D = \{D^c, D^{m_1}, D^{m_2}\}$, where $D^c = \{x_i^{m_1}, x_i^{m_2}, y_i\}$ represents the subset with complete pairs, and $D^{m_1} = \{x_j^{m_1}, y_j\}$ and $D^{m_2} = \{x_k^{m_2}, y_k\}$ denote incomplete modalities (i.e., only m_1 or m_2 is available). To align with the input requirements of the multimodal model, we use dummy inputs to simulate missing modalities (e.g., empty characters/pixels for text/image). Specifically, the subsets D^{m_1} and D^{m_2} are transformed into $\tilde{D}^{m_1} = \{x_j^{m_1}, \tilde{x}_j^{m_2}, y_j\}$ and $\tilde{D}^{m_2} = \{\tilde{x}_k^{m_1}, x_k^{m_2}, y_k\}$, respectively, to ensure consistent input structure for the model. $\tilde{x}^{m_1}$ and $\tilde{x}^{m_2}$ denote dummy inputs for the missing modalities. The resulting training dataset, comprising both complete data and transformed incomplete data, is defined as $\tilde{D} = \{D^c, \tilde{D}^{m_1}, \tilde{D}^{m_2}\}$.

3.2 The Overall Framework

Given the pretrained multimodal model F, which comprises a patch embedding layer and stacked transformer blocks, each block consists of a multi-head self-attention (MHSA) and an MLP layer. For multimodal input, each modality is encoded separately and combined with position embeddings, modality-type embeddings, and class embeddings to produce $f_{m_1}^0 \in \mathbb{R}^{L_{m_1} \times D}$ and $f_{m_2}^0 \in \mathbb{R}^{L_{m_2} \times D}$, where L_m represents the embedding length and D denotes the embedding dimension. The encoded representations are then concatenated to form the model input: $h^0 = [f_{m_1}^0, f_{m_2}^0]$. Each transformer block is defined as Eq.(1):

$$h^{i+1} = f_\theta^i(h^i) \qquad i \in [0, l) \tag{1}$$

where f_θ^i represents the parameters of the i-th block, and l denotes the total number of transformer blocks.

To address missing modality scenarios, we propose Memory-Driven Prompt Learning (**MemPrompt**), which compensates for missing modality information by querying semantically similar samples within the current modality and leveraging cross-modal associations. The whole framework is shown in Fig.1. Specifically, these two compensation strategies are implemented using generative prompts and shared prompts, respectively. The construction process of these prompts is detailed in Sec. 3.3. Here, we define the generative prompts as $P_{g,m}$ and the shared prompts as $P_{s,m}$, where $m = \{m_1, m_2\}$. Unlike previous approaches that directly concatenate prompts with input representations, our method constructs separate prompts for each modality and prepends them to their respective modality features. After adding the prompts, the input to each transformer block is represented as:

$$h^i = [p_{s,m_1}^i, p_{g,m_1}^i, f_{m_1}^i, p_{s,m_2}^i, p_{g,m_2}^i, f_{m_2}^i] \tag{2}$$

where $f_{m_1}^i$ and $f_{m_2}^i$ denote the features of each modality in the i-th transformer block. The newly constructed h^i is then processed by Eq.(1) to produce the final outputs.

During training, the pretrained backbone remains frozen, and only the added prompts and task-specific layers are fine-tuned. Let θ represent the additional prompts-related parameters and θ_t represent the task-specific parameters of the classification layer. The final loss function is defined as:

$$L = L_{task}(x_{m_1}^i, x_{m_2}^i; \theta, \theta_t) \tag{3}$$

where $x_{m_1}^i$ and $x_{m_2}^i$ denote inputs that may contain missing modalities, and L_{task} represents the task-specific training loss, typically the cross-entropy loss.

3.3 Prompts Design

In unified transformer models, disentangling individual modalities from the unified representation is challenging. To address this, we employ a prompt-based strategy to compensate for missing information. Specifically, the compensation is derived from two sources: (1) semantically similar samples within the current modality and (2) complementary information from the available modalities. To achieve

these goals, we decompose the vanilla prompts into generative prompts and shared prompts. The generative prompts are adaptively constructed from a predefined prompt memory based on the current input. This prompt memory stores prior information about specific modalities, enabling it to provide approximate feature representations to compensate for missing modalities. In contrast, shared prompts primarily establish cross-modal relationships. When one modality is missing, the shared prompts allow the available modality to provide complementary information. Unlike previous approaches, our method does not rely on predefined missing-modality scenarios, enabling it to effectively handle a wider range of missing-modality conditions. Details of the generative prompts and shared prompts are discussed below.

Generative Prompts

Unlike custom-designed prompts tailored to specific missing-modality cases, generative prompts are adaptively constructed based on the current input, providing a flexible mechanism to enhance the robustness of multimodal models. Specifically, let the input features at the i-th transformer block be $h^i = [f^i_{m_1}, f^i_{m_2}]$, where $f^i_{m_1}$ and $f^i_{m_2}$ represent the features of each modality. For each modality, the generative prompts are constructed as:

$$p^i_{g,m} = \text{PromptMem}(f^i_m) \qquad m \in \{m_1, m_2\} \tag{4}$$

where PromptMem is the prompt memory that contains a set of prompts and their associated weights, storing prior feature information for each modality to effectively handle missing-modality scenarios. During training, the feature information of each modality is stored in the memory. When a sample encounters a missing modality, similar features are retrieved from the memory, and generative prompts are adaptively built to compensate for the missing information.

The prompt memory is defined as $\text{PromptMem} = \{(p_k, q_k), k \in [1, K]\}$, where $p_k \in \mathbb{R}^{L\times D}$ represents the prompt embeddings, $q_k \in \mathbb{R}^D$denotes the associated coefficients, K is the memory size, L is the prompt length, and D is the feature dimension. To evaluate the similarity between the current input and the prompts in the memory, we first compute the cosine similarity between f_m and q_k, followed by softmax normalization:

$$w_k = \text{softmax}(\cos(f_m, q_k)) \tag{5}$$

where w_k represents the similarity between the current input and each prompt in the memory. The N most similar prompts are then selected, and a weighted summation is performed to generate the final generative prompt:

$$p_{g,m} = \sum_{i=1}^{N} w_k \cdot p_k \tag{6}$$

Shared Prompts

In multimodal models, inter-modal semantic associations exist between different modalities and can be leveraged to compensate for missing modality information. Specifically, when one modality is absent, the available modalities can utilize these associations to partially compensate for the missing information. To achieve this, we establish cross-modal connections by shared prompts, thereby enhancing the model's robustness under missing-modality scenarios.

Shared prompts, denoted as p^i_s, are first constructed and then mapped to the respective modalities. This design allows the shared prompts to capture common features across modalities while the mapping functions learn modality-specific information. The mapping process is defined as:

$$p^i_{s,m} = \mathbb{G}_m(p^i_s) \qquad m \in \{m_1, m_2\} \tag{7}$$

where $\mathbb{G}_m$ is a bottleneck network consisting of an MLP and activation functions, significantly reducing the trainable parameters.

The constructed generative prompts $p_{g,m}$ and shared prompts $p_{s,m}$ are integrated to Eq.(2). The updated h^i are send to Eq.(1) to derive the output for each transformer block.

4 Experiments

4.1 Datasets

We evaluated our proposed method on three widely used datasets: MM-IMDb [Arevalo *et al.*, 2017], UPMC Food101 [Wang *et al.*, 2015], and Hateful Memes [Kiela *et al.*, 2020].

MM-IMDb. The MM-IMDb dataset is a benchmark for multimodal movie genre classification, consisting of 25,959 movies. Each sample includes textual and visual information, such as movie posters and plot summaries, making it a challenging dataset for integrating diverse modalities.

UPMC Food101. The UPMC Food101 provides a multimodal benchmark for food image classification, featuring images and rich textual metadata across 101 food categories. The metadata includes ingredient lists and cooking methods, offering critical contextual information to enhance model performance in understanding complex visual and textual cues.

Hateful Memes. The Hateful Memes dataset, developed by Facebook AI, is designed to evaluate multimodal hate speech detection. This datasets is designed to promote multimodal models perform well while restrain the unimodal models performance by adding challenging samples ("benign confounders").

4.2 Implementation Details

For fair comparison, we adopted the same data processing methods as in prior work on the three benchmarks [Lee *et al.*, 2023]. The maximum text input lengths were set according to the specific datasets: 1024 for MM-IMDb, 512 for UPMC Food-101, and 128 for Hateful Memes. We employed the multimodal transformer ViLT [Kim *et al.*, 2021] as the backbone model. The pretrained ViLT model was frozen, and only the learnable prompts and classification layer were fine-tuned on the target datasets. The lengths of the generative and shared prompts were set to 16 for MM-IMDb and UPMC Food101, and 4 for Hateful Memes. This is due to the short text input length (128 tokens) of Hateful Memes necessitated shorter prompt lengths to facilitate model optimization. The prompt memory was configured with N as 5 and memory size as 16. Generative and shared prompts were added only to the first 6 transformer blocks. All experiments used the AdamW optimizer, with an initial learning rate of 5×10^{-3} for MM-IMDb and UPMC Food-101, and 1×10^{-3} for Hateful Memes. The weight decay rate was set to 2×10^{-2}. All experiments were conducted on NVIDIA 4090Ti GPUs.

Dataset	Missing Rate	Training Image	Training Text	Inference Image	Inference Text	ViLT	MAP (Attention)	MAP (Input)	Ours
MM-IMDb (F1-Macro)	70%	100%	30%	100%	30%	35.09	35.82	36.03	**38.47**
				30%	100%	28.21	28.76	21.02	**35.06**
				65%	65%	31.77	33.33	30.72	**37.60**
		30%	100%	100%	30%	29.64	25.30	22.08	**34.86**
				30%	100%	37.65	44.46	44.64	**48.59**
				65%	65%	34.71	36.38	35.00	**42.14**
		65%	65%	100%	30%	34.23	35.77	37.55	**39.09**
				30%	100%	35.19	42.08	44.35	**45.37**
				65%	65%	36.10	39.51	41.48	**42.47**
Food101 (Accuracy)	70%	100%	30%	100%	30%	67.25	73.54	74.25	**75.15**
				30%	100%	46.13	27.82	27.82	**80.81**
				65%	65%	51.90	51.38	51.17	**78.11**
		30%	100%	100%	30%	42.38	39.98	29.68	**60.63**
				30%	100%	76.43	85.68	86.30	**86.52**
				65%	65%	58.90	62.68	57.93	**74.07**
		65%	65%	100%	30%	64.47	71.31	72.24	**72.39**
				30%	100%	73.04	85.28	85.91	**86.12**
				65%	65%	70.13	78.39	79.11	**79.76**
Hateful Memes (AUROC)	70%	100%	30%	100%	30%	60.95	**61.66**	59.13	60.26
				30%	100%	59.36	59.44	57.55	**59.64**
				65%	65%	61.37	58.50	58.08	**62.55**
		30%	100%	100%	30%	56.07	59.04	59.32	**59.82**
				30%	100%	63.11	64.57	65.27	**65.63**
				65%	65%	59.42	58.33	58.82	**65.12**
		65%	65%	100%	30%	59.48	61.06	61.03	**61.91**
				30%	100%	62.91	61.75	64.05	**64.26**
				65%	65%	61.80	58.10	60.35	**65.74**

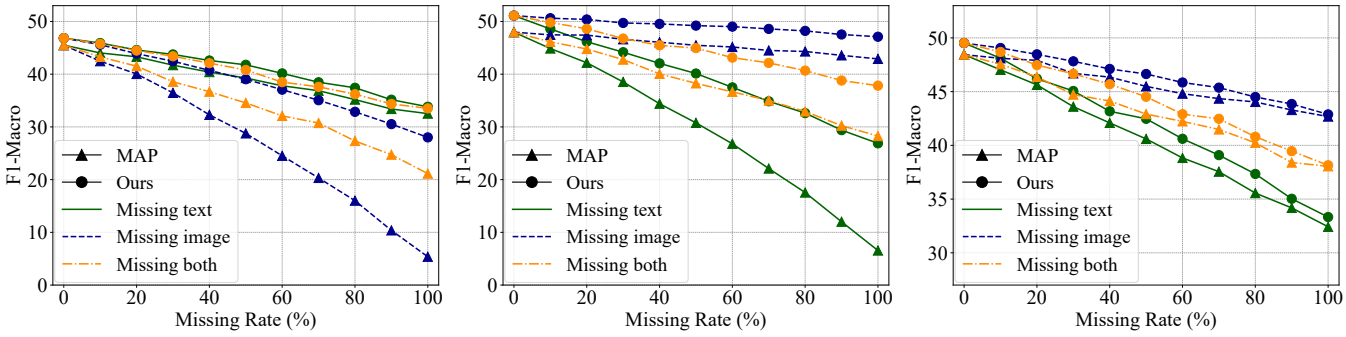
Table 1: Quantitative results on the MM-IMDB, UPMC Food-101, and Hateful Memes benchmarks. The missing rate $\eta\%$ is set to 70% during both train and inference. Given the fix missing case during training, we evaluate all models under various missing cases during inference. The best performance is in bold.

Missing modalities setting. During training, we followed the missing scenarios configurations outlined in [Lee *et al.*, 2023]. Three missing-modality scenarios were examined: missing text, missing image, and missing both. For the missing text and missing image cases, a total missing ratio of $\eta\%$ was applied, while for the missing both scenario, each modality was missing at a ratio of $\frac{\eta}{2}\%$ individually. The missing data were simulated by replacing the corresponding modality with dummy inputs (e.g., empty characters for text and blank pixels for images). To evaluate the generalization of the model, we trained it with specific missing modality configurations and tested it across all three scenarios.

4.3 Results and Analysis

Various missing modality scenarios are configured during both training and inference, including inconsistent missing cases between these stages. This setup provides a comprehensive evaluation of the model's robustness and generalization. Given a specific missing modality scenario during training, we assessed the model performance under different missing scenarios during inference. We compared our proposed method with two existing approaches, ViLT [Ma *et al.*, 2022] and MAP [Lee *et al.*, 2023], across three benchmarks: MM-IMDb, Food101, and Hateful Memes. The experimental results are summarized in Tab.1. Attention and input refer to different prompt placement strategies within MAP.

The compared results reveal two key points: First, when the missing modality scenario during training and inference are consistent, our model demonstrates noticeable improvements over the previous SOTA methods. For instance, on the MM-IMDb dataset, our model outperforms previous method under different missing conditions, achieving improvements such as 36.03 to 38.47 in the text-missing case, 44.64 to 48.59 in the image-missing case, and 41.48 to 42.47 in the both-missing case. Second, when the missing modality scenarios during training and inference are inconsistent, our model significantly outperforms the baseline methods. For instance, on the Food-101 dataset, when the model is trained under the text missing scenario but tested under the image missing scenario, our model achieves a remarkable performance improvement from 27.82 to 80.81. Conversely, when trained with missing images and tested with missing text, the performance increases from 39.98 to 60.63. These results demonstrate that our model effectively compensates for missing modality information, enabling it to handle unseen missing modality scenarios during inference.



(a) Training: Missing Text 70%. (b) Training: Missing Image 70%. (c) Training: Missing Both 70%.

Figure 2: Quantitative results across various missing rates on the MM-IMDb dataset. We set the missing cases as missing text, missing image, and missing both respectively during training, with a fixed missing ratio of 70%. During inference, the model is evaluated under different missing cases across varying missing rates.

Missing type	Baseline	w/ p_s	w/ p_g	Full
Missing Text 70%	31.69	32.86	35.13	37.04
Missing Image 70%	34.00	39.73	40.81	41.86
Missing Both 70%	35.17	40.74	41.25	42.31

Table 2: Ablation studies. We evaluate the effectiveness of each module on MM-IMDb datasets. p_s indicates shared prompts, while p_g represents generative prompts. we configured missing text, missing image, and missing both scenarios with 70% missing ratio during training. The results represent the average performance across all missing scenarios.

4.4 Ablation Studies

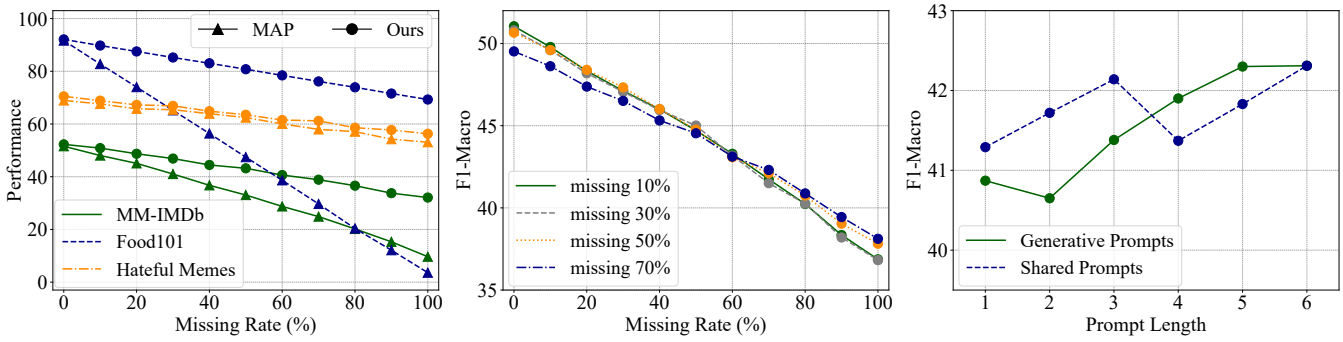
Evaluation under different missing rates. This section evaluates the model's robustness under varying missing rates. The missing rate is fixed during training, while the model is assessed under various missing rates during inference. Experiments were conducted on the MM-IMDb dataset. During training, missing cases were configured as follows: 70% missing text, 70% missing images, and 70% missing both. For inference, the missing rates ranged from 0% to 100%. The results are presented in Fig. 2. The results indicate that our method consistently outperforms previous models across all scenarios. When the missing cases during training and inference are consistent, our method achieves slightly better performance. Notably, when the missing cases during training and inference are inconsistent, previous models exhibit a significant performance drop, while our model remains stable across varying missing settings.

Furthermore, as the missing ratio increases, the performance of our model degrades more gradually, demonstrating its ability to effectively compensate for missing information. These findings emphasize the adaptability of our method under varying missing ratios, highlighting its superior capability to maintain performance across unseen missing configurations compared to previous methods.

Effect of each module. Our method incorporates two key components: generative prompts and shared prompts. To evaluate the contribution of each module, we conduct ablation studies on the MM-IMDb dataset, assessing the model performance under three different missing cases. Using the ViLT pretrained model as the baseline, we separately add generative prompts and shared prompts to examine their individual impact. Results show that adding either generative prompts or shared prompts alone significantly improves the model's robustness in handling missing modality scenarios. Furthermore, the last column in Tab. 2, representing the full model with both prompts, demonstrates that combining generative and shared prompts provides a synergistic effect, achieving the highest performance across all settings. This confirms that the two modules complement each other, with generative prompts dynamically retrieving semantically similar features and shared prompts establishing cross-modal associations to compensate for missing modalities more comprehensively.

Comparison with complete training data. In real-world scenarios, complete data pairs are sometimes available for training. To further evaluate our model, we trained it on complete data and evaluated its performance under various missing rates during inference. For evaluation, we averaged the results across the missing-text, missing-image, and missing-both cases to derive the final metric. The results, shown in Fig. 3(a), indicate that all models perform well when the test data is also complete. However, as the missing rate increases, the performance of previous methods declines significantly, while our model demonstrates greater stability and robustness.

Results of different missing rates during training. We evaluated the model by training with different missing rates (10%, 30%, 70%, 90%) and testing under varying missing rates during inference, with results shown in Fig. 3(b). In this part, the missing-both case was used for training, and the results represent the average performance across the missing-image, missing-text, and missing-both cases. Our model demonstrates strong generalization ability, maintaining consistent performance across different missing rates. This suggests that the proposed method effectively compensates for missing information, allowing the model to adapt to diverse scenarios. Notably, under high missing rates during training, the model achieves better performance when tested under



(a) Training with complete data (b) Training under different missing ratio (c) Effect of prompt layers

Figure 3: More ablation studies. The results are averaged over different missing scenarios (missing text, missing image, and missing both). (a) Model trained on complete data and tested under various missing scenarios across different missing rates. (b) Model trained with different missing rates (10%, 30%,50%,70%) under missing both case, with results averaged over all missing scenarios. (c) Effect of prompt layers: Generative prompts are fixed in first 6 layers while varying the shared prompt layers from 1 to 6, and vice versa.

Memory Size	8	12	16	20	24
Missing Text	35.99	36.68	37.04	35.41	35.43
Missing Image	41.47	40.75	41.86	41.27	41.57
Missing Both	41.67	41.51	42.31	41.08	41.52

Table 3: Impact of prompt memory size.

Prompt Location	Pre-Append	Distributed
Missing Text	33.98	37.04
Missing Image	40.74	41.86
Missing Both	41.51	42.31

Table 4: Results of different prompt location.

similarly high missing rates. This result aligns with previous findings [Lee *et al.*, 2023] and highlights the robustness of our approach in handling incomplete data distributions.

Number of prompt layers. In the proposed method, we designed generative prompts and shared prompts and integrated them into the pretrained model. This section examines the impact of the number of prompt layers on model performance. Specifically, generative prompts are fixed in 6 layers, while the number of shared prompt layers is varied from 1 to 6. Conversely, shared prompts are fixed in 6 layers, and the number of generative prompt layers is varied. The results indicate that although performance fluctuates slightly at certain points, the overall trend shows improvement as the number of layers increases.

Analysis of memory size. The prompt memory is designed for each modality to store its feature information through a series of prompts. Generative prompts are adaptively constructed by retrieving relevant information from the prompt memory to compensate for missing modalities. In this section, we analyze the impact of memory size K on model performance using the MM-IMDb dataset. As shown in Tab. 3, model performance initially improves as memory size increases, reaching its peak at $K = 16$. This suggests that a larger memory provides more diverse prompts and effectively captures modality-specific features. However, further increasing the memory size slightly reduces performance, likely due to increased complexity. These results highlight the importance of selecting an appropriate memory size to balance adaptability and efficiency.

Effect of prompt location. In our proposed method, prompts for each modality are directly added to their associated modality features, as shown in Eq. (2). To analyze the impact of prompt location, we conducted an ablation study comparing two strategies: Pre-Appended Prompting and Distributed Prompting. The Distributed Prompting strategy, as defined in Eq. (2), places generative and shared prompts directly before their respective modality features. In contrast, the Pre-Appended Prompting strategy combines the generative and shared prompts of all modalities and appends them at the beginning of the input sequence. The results, shown in Tab. 4, demonstrate that Distributed Prompting consistently outperforms Pre-Appended Prompting. This suggests that aligning prompts closely with their associated features leads to better model performance.

5 Conclusion

In this paper, we proposed Memory-Driven Prompt Learning, a novel framework designed to tackle the challenges posed by missing modalities in multimodal learning. Our approach introduces a compensation mechanism that integrates two complementary strategies: retrieving semantically relevant information from the missing modality itself and leveraging cross-modal associations derived from the available modalities. These strategies are realized through two types of prompts—generative prompts, which dynamically retrieve information from a prompt memory to compensate for missing modality features, and shared prompts, which infer and supplement common semantic information from the available modalities. Extensive experiments on multiple benchmarks demonstrate that our method significantly enhances model robustness and achieves state-of-the-art performance under diverse missing modality scenarios.

Acknowledgments

This work was supported by the National Key R&D Program of China under grant 2022YFF0901800, and the National Natural Science Foundation of China (NSFC) under grants No. 62176205, 62302383, and 62372365.

References

[Akbari *et al.*, 2021] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

[Arevalo *et al.*, 2017] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[Bao *et al.*, 2022] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[Han *et al.*, 2023] Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, and Dongfang Liu. E 2 vpt: An effective and efficient approach for visual prompt tuning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17445–17456. IEEE, 2023.

[He *et al.*, 2023a] Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. Sensitivity-aware visual parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11825–11835, 2023.

[He *et al.*, 2023b] Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. Parameter-efficient model adaptation for vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 817–825, 2023.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

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

[Khattak *et al.*, 2023b] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023.

[Kiela *et al.*, 2020] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

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

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. 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, 2021.

[Li *et al.*, 2024] Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *arXiv preprint arXiv:2404.15159*, 2024.

[Ma *et al.*, 2021] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021.

[Ma *et al.*, 2022] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.

[Nagrani *et al.*, 2021] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

[Park *et al.*, 2024] Jinyoung Park, Juyeon Ko, and Hyunwoo J Kim. Prompt learning via meta-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26940–26950, 2024.

[Petrov *et al.*, 2024] Aleksandar Petrov, Philip HS Torr, and Adel Bibi. Prompting a pretrained transformer can be a universal approximator. *arXiv preprint arXiv:2402.14753*, 2024.

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

[Reza *et al.*, 2024] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Roy and Etemad, 2024] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[Shao *et al.*, 2023] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 14974–14983, 2023.

[Shi *et al.*, 2024] Zhiyi Shi, Junsik Kim, Wanhua Li, Yicong Li, and Hanspeter Pfister. Mora: Lora guided multi-modal disease diagnosis with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–282. Springer, 2024.

[Stefanini *et al.*, 2022] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022.

[Wang *et al.*, 2015] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015.

[Wang *et al.*, 2016] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[Wang *et al.*, 2023] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023.

[Wang *et al.*, 2024] Haixin Wang, Jianlong Chang, Yihang Zhai, Xiao Luo, Jinan Sun, Zhouchen Lin, and Qi Tian. Lion: Implicit vision prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5372–5380, 2024.

[Woo *et al.*, 2023] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2776–2784, 2023.

[Wu *et al.*, 2024a] Haoyuan Wu, Haisheng Zheng, Zhuolun He, and Bei Yu. Parameter-efficient sparsity crafting from dense to mixture-of-experts for instruction tuning on general tasks. *arXiv preprint arXiv:2401.02731*, 2024.

[Wu *et al.*, 2024b] Renjie Wu, Hu Wang, Feras Dayoub, and Hsiang-Ting Chen. Segment beyond view: Handling partially missing modality for audio-visual semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6100–6108, 2024.

[Xin *et al.*, 2024a] Yi Xin, Junlong Du, Qiang Wang, Zhiwen Lin, and Ke Yan. Vmt-adapter: Parameter-efficient transfer learning for multi-task dense scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16085–16093, 2024.

[Xin *et al.*, 2024b] Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16076–16084, 2024.

[Yang *et al.*, 2024] Zequn Yang, Yake Wei, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. *arXiv preprint arXiv:2402.06244*, 2024.

[Zhao *et al.*, 2024a] Yihan Zhao, Wei Xi, Yuhang Cui, Gairui Bai, Xinhui Liu, and Jizhong Zhao. Copl: Parameter-efficient collaborative prompt learning for audio-visual tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4455–4464, 2024.

[Zhao *et al.*, 2024b] Zhida Zhao, Jia Li, Lijun Wang, Yifan Wang, and Huchuan Lu. Maskmentor: Unlocking the potential of masked self-teaching for missing modality rgb-d semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1915–1923, 2024.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.