MuAP: Multi-step Adaptive Prompt Learning for Vision-Language Model with Missing Modality

Anonymous ACL submission

Abstract

Recently, prompt learning has garnered considerable attention for its success in various Vision-Language (VL) tasks. However, existing prompt-based models are primarily focused on studying prompt generation and prompt strategies with complete modality settings, which does not accurately reflect real-world scenarios where partial modality information may be missing. In this paper, we present the first comprehensive investigation into prompt learning behavior when modalities are incomplete, revealing the high sensitivity of prompt-based models to missing modalities. To this end, we propose a novel **Mu**lti-step **A**daptive **P**rompt Learning (**MuAP**) framework, aiming to generate multimodal prompts and perform multi-step prompt tuning, which adaptively learns knowledge by iteratively aligning modalities. Specifically, we generate multimodal prompts for each modality and devise prompt strategies to integrate them into the Transformer model. Subsequently, we sequentially perform prompt tuning from single-stage and alignment-stage, allowing each modality-prompt to be autonomously and adaptively learned, thereby mitigating the imbalance issue caused by only textual prompts that are learnable in previous works. Extensive experiments demonstrate the effectiveness of our MuAP and this model achieves significant improvements compared to the state-of-the-art on all benchmark datasets. Our codes are available at https://anonymous.4open.science/r/multiview_adaptative_prompt_learning/.

1 Introduction

Vision-Language (VL) pre-training (Su et al., 2019; Lu et al., 2019; Yu et al., 2019; Kim et al., 2021) has demonstrated remarkable success in various Vision-Language tasks like image recognition (Zhang et al., 2021; Liu et al., 2019), object detection (Jin et al., 2021; Sun et al., 2021), and image segmentation (Cao et al., 2021; Hu et al., 2019) by learning the semantic correlations between different modalities through large-scale image-text training. However, most previous research has assumed that all modalities are accessible during both training and testing phases, a condition that is often challenging to meet in real-world scenarios. This challenge arises from various factors, such as privacy and security concerns leading to the inaccessibility of textual data (Lian et al., 2023), or limitations in device observations resulting in missing visual data (Zeng et al., 2022; Ma et al.,

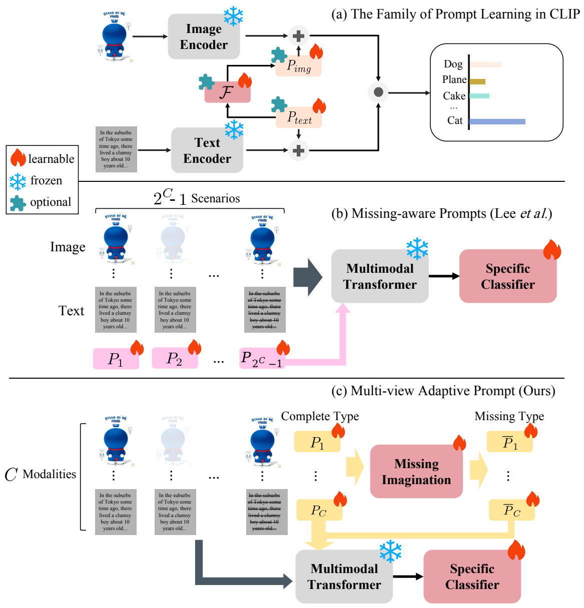


Figure 1: Various architectures in the prompt tuning field. (a) The CLIP-family method (Khattak et al., 2023) focus on prompt generation with complete modality information. (b) Missing-aware prompts method in MPVR (Lee et al., 2023) has $2^C - 1$ prompts to represent all missing scenarios, where C is the number of modalities. (c) Our method aims to enhance parameter efficiency by utilizing only C prompts and to improve robustness through multi-step prompting tuning in missing scenarios.

2022). Hence, the widespread occurrence of missing modalities distinctly hinders the performance of vision-language models.

Recently, as shown in Figure 1(a), there has been a notable advancement in the field of visual language (VL) by adopting prompt learning from Natural Language Processing (NLP). However, researchers do not consider scenarios where modalities are missing. For instance, CLIP (Radford et al., 2021) aligns image and language modalities through joint training on large-scale datasets. It leverages handcrafted prompts and a parameterized text encoder to generate precise classification weights, thereby enabling zero-shot learning. Nonetheless, it faces two formidable challenges: the need for expertise and multiple iterations in designing handcrafted prompts, as well as the impracticality of fully fine-tuning the entire model due to its tremendous scale. Consequently, CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) propose automated prompt engineering that converts contextual words in prompts into learnable vectors and achieves substantial improvements by exclusively fine-tuning dense prompts using a small number of labeled images. Furthermore, MaPLe (Khattak et al., 2023) delves into the limitations of solely using language prompts in previous works and presents multimodal prompt learning, which introduces a coupling function to connect text prompts with image prompts, facilitating mutual gradient propagation between the two modalities for more precise alignment.

Recent research, such as MPVR (Lee et al., 2023), has proposed using prompt learning for scenarios with missing modalities, aiming to mitigate the performance degradation caused by disparities in modality absence in training or testing data samples. However, designing distinct prompts for each missing modality scenario inevitably leads to an exponential increase in the number of prompts as the number of modalities increases (as shown in Figure 1(b), a scenario with C modalities necessitates $2^C - 1$ prompts), seriously compromising the scalability of the model. Moreover, unlike the dual-prompt strategy used by MaPLe (Khattak et al., 2023), MPVR (Lee et al., 2023) adopts a coarse prompt strategy at the input or attention level by directly inserting prompts into multimodal transformers, without distinguishing textual and visual features.

Despite MaPLe 's (Khattak et al., 2023) dual-prompt strategy effectively harnessing the capabilities of both modalities, its coupling mechanism exhibits a propensity for relying predominantly on the textual modality, which may result in unbalanced learning of multimodal information. Furthermore, an excessive degree of coupling has the potential to impede the independent learning capacity of each modality. To address this, in Figure 1(c), we propose a novel **<u>Mu</u>**lti-step **<u>A</u>**daptative **<u>P</u>**rompt Learning (**MuAP**) framework for multimodal learning in the presence of missing modalities. MuAP introduces a multi-step prompting mechanism that adaptively learns multimodal prompts by iteratively aligning modalities. Specifically, we perform prompt tuning sequentially from two perspectives: single-stage and alignment-stage. This allows each modality prompt to learn autonomously without interference from the other, facilitating an in-depth exploration of each modality in scenarios where certain modalities are missing. Finally, we obtain the downstream classifier results through multimodal prompt learning, where adaptive prompts effectively mitigate imbalanced learning caused by one-way coupling and only textual prompts are learnable in (Khattak et al., 2023).

To summarize, this paper makes the following key contributions:

- To the best of our knowledge, this paper is the first study to analyze the robustness of prompt learning on missing modality data. We propose a novel missing-modality in the VL Model model with multi-step adaptive prompt learning, addressing the limitations of previous works and enhancing prompts through autonomous and collaborative learning simultaneously.

- We devise a multi-step tuning strategy that encompasses single-stage and alignment-stage tunings, where we generate visual and language prompts adaptively through multi-step modality alignments for multimodal reasoning. This facilitates comprehensive knowledge learning from both modalities in an unbiased manner.

- We conduct extensive experiments and ablation studies on three benchmark datasets. Extensive experiments demonstrate the effectiveness of our MuAP and this model achieves significant improvements compared to the state-of-the-art on all benchmark datasets.

2 Related work

2.1 Vision-Language Pre-trained Model

Recent researches on Vision-Language Pre-training (VLP) aim to learn semantic alignment between different modalities by leveraging large-scale image-text pairs. There are two architectures of the existing VLP methods: single-stream and dual-stream architectures. In single-stream architectures, image and text representations are concatenated at the feature level and serve as input to a single-stream Transformer. For example, VisualBERT (Li et al., 2019) concatenates text embedding sequences and image embedding sequences, which are then passed through a Transformer network. Building upon this work. VL-BERT (Su et al., 2019) utilizes OD-based Region Features on the image side and incorporates a Visual Feature Embedding module. Similarly, ImageBERT (Qi et al., 2020) follows a single-stream model with OD for image feature extraction while introducing more weakly supervised data to enhance learning performance. Alternatively, the dual-stream architectures align image-text representations in a high-level semantic space using two separate cross-modal Transformers. For instance, CLIP (Radford et al., 2021) and its variants (such as CoOp (Zhou et al., 2022b) and MaPLe (Khattak et al., 2023)) employ ResNet (He et al., 2016) and ViT models as image encoders, while employing Transformers (Vaswani et al., 2017) as text encoders. Subsequently, they utilize contrastive learning to predict matching scores between each template entity and the current image, with the highest score indicating the image's classification result.

2.2 Prompt Learning for Vision-Language Tasks

As the diversity of Vision-Language (VL) tasks poses a challenge for individually fine-tuning large pre-trained models for each task, Prompt Learning emerges as an effective approach to tackle this challenge. It involves freezing the backbone neural network and introducing prompts, which comprise a small number of trainable parameters, to fine-tune the entire model. This allows for the zero-shot or few-shot application of pre-trained models to new VL tasks in a more parameter-efficient manner than training large models from scratch for each task. For example, CoOp (Zhou et al., 2022b) incorporates learnable prompts into the language encoder to fine-tune CLIP, while CoCoOp employs conditional prompts to further enhance the model's generalization ability. MaPLe (Khattak et al., 2023) argues that learning prompts for the text encoder alone in CLIP are insufficient to model the necessary adaptations required for the image encoder. To address this, MaPLe leverages multimodal prompt learning to fully fine-tune the text and image encoder representations, ensuring optimal alignment in downstream tasks. It employs a coupling function to connect the prompts learned in the text and image encoders, with only the text prompts being trainable.

3 Method

In this section, we detail our methodology by presenting a clear problem definition and introducing our proposed MuAP.

3.1 Problem Definition

In this work, we study the missing-modality multimodal learning where the presence of missing modalities can occur in both the training and testing phases. For simplicity while retaining generality, following (Huang et al., 2019), we consider a multimodal dataset that contains two modalities: $\mathcal{M} = \{m_t, m_v\}$, where m_t and m_v denote textual, visual modalities respectively. The complete modality data can be represented as $\mathcal{R}^{all} = \{x_i^{m_t}, x_i^{m_v}, y_i\}$, where $x_i^{m_t}$ and $x_i^{m_v}$ denote the textual and visual features respectively, y_i denotes the corresponding class label. While the missing modality data are $\mathcal{R}^{m_t} = \{x_j^{m_t}, y_j\}$ or $\mathcal{R}^{m_v} = \{x_k^{m_v}, y_k\}$ representing text-only data and image-only data respectively. To keep the format of multimodal inputs, we adopt a straightforward strategy of assigning placeholder inputs, represented as $\overline{x}^{m_t}$ and $\overline{x}^{m_v}$, to the instances with missing modalities. These placeholder inputs are null strings or blank pixels and serve to fill the absence of textual or visual data, respectively. Consequently, we obtain $\overline{\mathcal{R}}^{m_t} = \{x_j^{m_t}, \overline{x}_j^{m_v}, y_j\}$, $\overline{\mathcal{R}}^{m_v} = \{\overline{x}_k^{m_t}, x_k^{m_v}, y_k\}$, and the multimodal data with missing modality can be represented as $\mathcal{R} = \{\mathcal{R}^{all}, \overline{\mathcal{R}}^{m_t}, \overline{\mathcal{R}}^{m_v}\}$. Our goal is to address classification issues and improve the robustness of VL model with Prompt Learning with missing modalities $\mathcal{R}$.

3.2 Overall Framework

Considering the resource constraints, we focus on the VL model with Prompt Learning and adopt Vision-and-Language Transformer (ViLT) (Kim et al., 2021) as the backbone, which is pre-trained

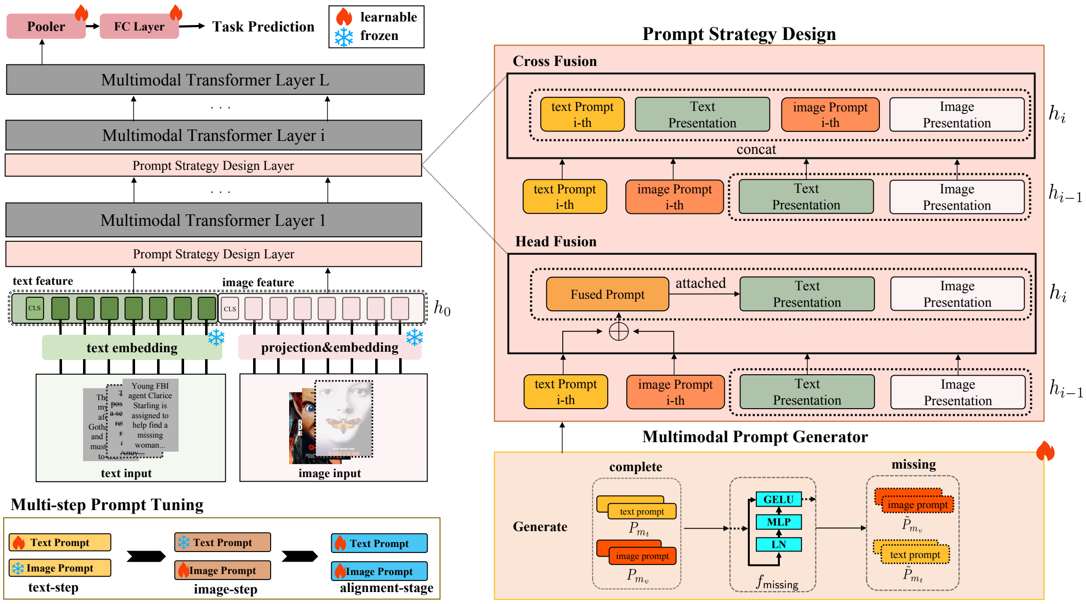


Figure 2: The overview of our MuAP framework. The Multimodal Prompt Generator initially generates complete-type prompts, P_{m_t} and P_{m_v}, tailored to the specific modality case (e.g., textual or visual modalities in Vision-Language tasks). Next, it employs f_{missing} to create missing-type prompts $\tilde{P}_{m_t}$ and $\tilde{P}_{m_v}$. The Prompt Strategy Design module integrates prompts into multiple MSA layers using various strategies (i.e., head fusion or cross fusion). During the training phase, we leverage Multi-step Prompt Tuning to synchronize distinct characteristics of different modality prompts effectively.

on large-scale VL datasets and remains untrainable in downstream tasks. To mitigate the significant performance degradation of Prompt Learning models due to missing modality data, we propose a novel **Mu**lti-step **A**daptative **P**rompt Learning (MuAP) model to enhance the model's robustness in various missing scenarios. As illustrated in Figure 2, MuAP mainly comprises three modules: Multimodal Prompt Generator, Prompt Strategy Design, and multi-step Prompt Tuning. Specifically, we first generate learnable specific prompts for each modality to achieve completeness tuning in prompting, deviating from previous methods (Zhou et al., 2022b,a) where only textual prompts were learnable. Subsequently, we introduce two prompt fusion strategies: head-fusion and cross-fusion, attaching prompts to blocks of the multimodal transformer. Additionally, we propose a multi-step tuning strategy for dynamic language and vision prompt tuning through modality alignments, allowing MuAP to gain knowledge from both modalities.

3.3 Revisiting ViLT

ViLT is a widely used Transformer-based multimodal pretraining model. It partitions images into patches of varying sizes, which are projected and embedded to generate latent representations. This allows the unified processing of images and text with minimal parameters. Its overall workflow commences by concatenating the text representation (denoted as $t = [t_{cls}; t_1; \ldots; t_M]$) with the image patches (denoted as $v = [v_{cls}; v_1; \ldots; v_N]$). These concatenated representations are then fed into multiple Transformer layers for processing. Specifically:

$$h^0 = [t + t^{modal}; v + v^{modal}] \in R^{L_V \times d} \quad (1)$$

$$\hat{h^i} = \mathsf{MSA}(\mathsf{LN}(h^{i-1})) + h^{i-1}, \quad i = 1 \ldots L \quad (2)$$

$$h^i = \mathsf{MLP}(\mathsf{LN}(h^{\widehat{i-1}})) + \hat{h^i}, \quad i = 1 \ldots L \quad (3)$$

where, t and v represent the embeddings of text and images, respectively. They are combined with their respective modality type embeddings t^{modal} and v^{modal} to form the initial input h^0. L_V represents the length of the input sequence, while d denotes the dimension of the hidden states. The context vectors h undergo continuous updates through L layers of Transformer encoders, and the final output context sequence h^L is utilized for downstream tasks.

3.4 Multimodal Prompt Generator

One main challenge in addressing missing modality learning with prompt learning lies in the design of prompt, and all modality absence situations are exponential. Drawing on the effectiveness of complete prompts in multimodal learning, we generate specific prompts for each modality, with the key distinction being that all the textual and visual prompts are both learnable. Unlike (Lee et al., 2023), where missing-aware prompts are generated for each possible situation resulting in an exponential increase as the number of modalities grows, our method adopts a linear growth pattern for prompts that significantly reduces the number of parameters and model complexity. To improve understanding and compensation for missing modalities, we create a simple network to generate specific prompts for each modality, aiding exploration and use of implicit data.

Specifically, when the input comprises C modalities, there exist C complete-type prompts. In our VL tasks, given $C = 2$ modalities of images and texts, we initialize P_{m_t} and $P_{m_v} \in R^{L_p \times d}$ as textual and image prompts respectively, representing the complete modality, where L_p is the prompt length. Subsequently, the initial prompts are fed into a lightweight network f_{missing}, in a crosswise manner. This means that opposing prompts are used to generate prompts (e.g., using a complete-type prompt from the visual modality to generate a missing-type prompt for the textual modality). The goal of this process is to enhance perception and compensate for missing modalities. The formula for the generating process is as follows:

$$f^i_{\text{missing}}(P^i) = \mathsf{GELU}(\mathbf{W}^i \mathsf{LN}(P^i)) + P^i \quad (4)$$

$$\tilde{P}^i_{m_v} = f^i_{\text{missing}}(P^i_{m_t}) \quad (5)$$

$$\tilde{P}^i_{m_t} = f^i_{\text{missing}}(P^i_{m_v}) \quad (6)$$

where $\mathbf{W}^i$ represents the weight matrix specific to the i-th f_{missing} module in the i-th layer of MSA, LN refers to the layer normalization operation, GELU is the activation function, and adding the original prompts P^i represents the residual operation. The residual connection is present to retain the opposing modality information while the MLP is utilized to collect additional missing-specific features to provide more valuable supplementary for the missing input and facilitate multimodal fusion. In a more generalized form, let P_m $(m \in \mathcal{M})$ represent the complete-type prompt for modality m, and $\tilde{P}_m$ represent the missing-type prompt for the same modality. When modality m is missing, the missing-type prompt $\tilde{P}_m$ is utilized in the subsequent module. Otherwise, the complete-type prompt P_m is used.

3.5 Prompt Strategy Design

Designing prompt template and strategy is crucial for prompt-based learning. We focus on prompt strategy involving prompt configuration and placement. Two prompt strategies introduced in Figure 2: head-fusion prompting and cross-fusion prompting. Consistency in subsequent symbols assumed with complete input data for textual and visual modalities.

Head-fusion Prompting. One simple way to incorporate prompts is to add them at the start of input sequences for each layer. We use element-wise summation for combining multimodal prompts. P_{head} is expressed as:

$$P_{head} = P_{m_t} \oplus P_{m_v}, P \in R^{L_p \times d} \quad (7)$$

where $\oplus$ denotes the summation over prompts from each modality. Next, we concatenate P_{head} with the input sequence of texts and images at each layer. Similar to ViLT (Kim et al., 2021), the formula can be expressed as follows:

$$h^i = [P^i_{head}; t^i; v^i], \quad i = 0 \cdots N_p \quad (8)$$

where P^i_{head} denotes the head-fusion prompt of i-th layer, N_p represents the number of MSA layers in ViLT. With the concatenating P^i_{head} to the input sequences of the previous layer, the final output length increases to $(N_P L_P + L_V)$ in total. This allows the prompts for the current layer to interact with the prompt tokens inherited from previous layers, enabling the model to learn more effective instructions for prediction.

Cross-fusion Prompting. Motivated by (Khattak et al., 2023), another prompting approach is to insert modality-specific prompts into their corresponding modality inputs in a single-stream model. By doing this, we facilitate the interaction between modality-specific prompts and features. The cross-fusion prompting can be formalized as follows:

$$h^i = [P^i_{m_t}; t^i; P^i_{m_v}; v^i], \quad i = 0 \cdots N_p \quad (9)$$

where $P^i_{m_t}$, $P^i_{m_v}$ represent the modality-specific prompts for the textual and visual modalities, respectively, at the i-th layer. It is noteworthy that, unlike (Khattak et al., 2023) which only replaces few parameters from the input sequence from each layer, cross-fusion prompt strategy follows head-fusion to attach the prompts at each MSA layer. This results in an expanded final output length of

$(2N_P L_P + L_V)$. This improves the model's representation scale and training stability, but it encounters a significant increase in model length when both N_P and L_P are large. It also faces the potential risk of overlooking the information in the original input sequence. We discuss how the prompt length leads to overfitting in Section 4.5.

3.6 Multi-step Prompt Tuning

In this section, we introduce our proposed multi-step prompt tuning technique designed to adaptively learn multimodal prompts through multi-step sequential modality alignments. Specifically, we employ prompt tuning (Lester et al., 2021) of the pre-trained Transformer encoder to perform efficient parameter learning from multiple stages, including single-stage of each modality and a alignment-stage. This not only facilitates the acquisition of modality-specific information from individual visual and textual modalities but also captures the correlations between different modalities.

Single-stage prompt tuning. To fully account for the inherent differences between distinct modalities, we sequentially and separately freeze the two modality prompts to explore learnable prompts trained with contrastive learning. As illustrated in Figure 2, we iteratively train the learnable prompts in a step-wise manner. Initially, we optimize the textual prompts while keeping the visual prompts frozen, called text-step. Subsequently, we switch to optimizing the visual prompts while fixing the textual prompts, called image-step. This exclusive updating process enables the prompt tuning to capture modality-specific attributes respectively.

Specifically, in the two steps, we utilize the Kullback-Leibler (KL) divergence as $\mathcal{L}_{kl}$ to measure the distribution difference between text and visual prompts. Additionally, we incorporate $\mathcal{L}_{cls}$ as a classification loss to facilitate the fusion.

To mitigate overfitting issues caused by prompt engineering, we employ diverse combinations of parameters λ_t and λ_v in the two steps of prompt updating, which effectively preserves modality-specific information. The formulas are as follows:

$$\textbf{Text-step}: \mathcal{L}_{total}^{t} = \mathcal{L}_{cls} + \lambda_t \mathcal{L}_{kl}(P_{m_t}, P_{m_v}) \tag{10}$$

$$\textbf{Image-step}: \mathcal{L}_{total}^{v} = \mathcal{L}_{cls} + \lambda_v \mathcal{L}_{kl}(P_{m_t}, P_{m_v}) \tag{11}$$

During this separate training of modality prompts, the hyper-parameter λ is used to combine with the KL loss. Specifically, λ_t and λ_v are set to 0.4 for the text prompt training step and 0.3 for the image prompt training step, respectively. In the process of single-stage prompt tuning, the two prompts undergo simultaneous updates through several alignment steps, with the experimental setup setting the number of steps to 3.

Alignment-stage prompt tuning. To further adapt multimodal prompts and enhance the generalization capability of downstream tasks, we train the model again from a alignment stage. In this step, the visual and textual prompts are all trainable during the training. The overall training objective solely emphasizes the classification loss $\mathcal{L}_{cls}$, which is formulated as follows:

$$\textbf{Alignment-stage}: \mathcal{L}_{total} = \mathcal{L}_{cls} \tag{12}$$

4 Experiments

4.1 Datasets and Metrics

Datasets We follow the approach outlined in (Lee et al., 2023) to evaluate our methods across three multimodal downstream tasks:

- ***MM-IMDb*** (Arevalo et al., 2017) focuses on classifying movie genres using both images and text, handling cases where a movie fits into more than one genre.
- ***UPMC Food-101*** (Wang et al., 2015) is a multimodal classification dataset and comprises 5% noisy image-text paired data gathered from Google Image Search.
- ***Hateful Memes*** (Kiela et al., 2020) is a challenging dataset for identifying hate speech in memes through images and text. It has 10k tough samples to challenge unimodal models and favor multimodal models.

Metrics Given the distinct classification tasks addressed by these datasets, we employ appropriate metrics tailored to each dataset. Specifically, for MM-IMDb, we utilize F1-Macro as a measure of multi-label classification performance. For UPMC Food-101, the metric is classification accuracy. For Hateful Memes, we assess performance using the AUROC.

4.2 Baselines

Baselines To assess the effectiveness and robustness of our proposed method, we primarily compare it with the state-of-the-art models. These models include

Datasets	Missing rate ϵ	Training Image	Training Text	Testing Image	Testing Text	ViLT	MPVR (Input-level)	MPVR (Attention-level)	Visual BERT (Li et al., 2019)	Ma Model (Ma et al., 2022)	MuAP (Head Fusion)	MuAP (Cross Fusion)
MM-IMDb (F1-Macro)	70%	30%	100%	30%	100%	37.61	46.30	44.74	38.63	46.63	**47.21**	46.73
		65%	65%	65%	65%	36.30	42.41	41.56	37.23	41.28	42.57	**43.92**
		100%	30%	100%	30%	34.71	39.19	38.13	36.41	38.65	**41.37**	39.88
Food101 (Accuracy)	70%	30%	100%	30%	100%	76.93	86.09	85.89	77.41	86.38	**86.90**	86.59
		65%	65%	65%	65%	69.03	77.49	77.55	71.06	78.58	77.87	**78.95**
		100%	30%	100%	30%	66.29	73.85	72.47	67.78	73.41	**74.61**	74.60
Hateful Memes (AUROC)	70%	30%	100%	30%	100%	61.74	62.34	63.30	61.98	63.56	65.09	**66.83**
		65%	65%	65%	65%	62.83	63.53	62.56	63.05	64.41	**64.76**	62.68
		100%	30%	100%	30%	60.83	61.01	61.77	60.89	60.96	**62.08**	61.26

Table 1: Quantitative results on the MM-IMDB (Arevalo et al., 2017), UPMC Food-101 (Wang et al., 2015), and Hateful Memes (Kiela et al., 2020) with missing rate $\xi\% = 70\%$. The outcomes were analyzed under diverse missing-modality cases, with the best results highlighted in **bold** for clarity.

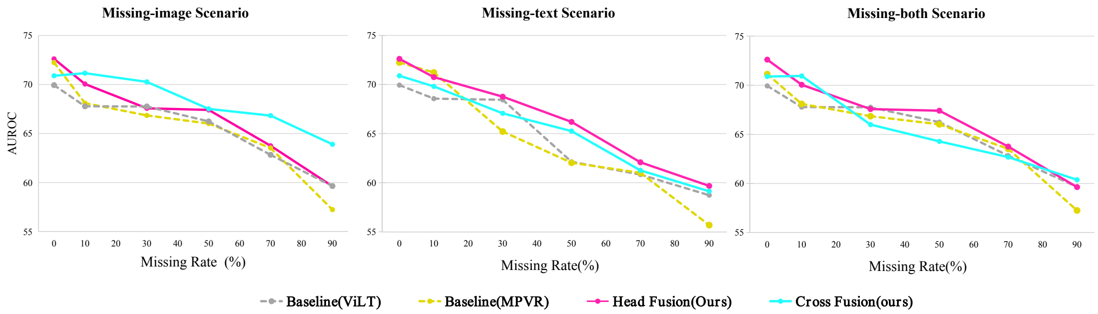


Figure 3: Comparison of baselines on the Hateful Memes dataset with different missing rates across various missing-modality scenarios. Each point in the picture represents training and testing with the same $\epsilon\%$ missing rate.

- ***Finetuned VILT***: the original one without any additional prompt parameters in ViLT (i.e. only training the pooler layer and task-specific classifier).

- ***MPVR (Lee et al., 2023)***: derived from the pre-trained VILT backbone, this model integrates **missing-aware prompts** into its multi-modal transformer design.

- ***Visual BERT (Li et al., 2019)***: a modified Visual BERT focusing on pooler and classifier training.

- ***Ma Model (Ma et al., 2022)***: using pre-trained VILT, multi-task optimization, and automated search algorithm to find most efficient fusion technique.

4.3 Main Results

Basic Performance. Table 1 shows our new prompt learning method outperforms baselines, demonstrating the effectiveness of our design and training strategy. The Hateful Memes dataset is tough, making unimodal models struggle, especially with missing modalities. Our head-fusion approach surpasses missing-aware prompts on this dataset, showing a 1.94% average improvement. This highlights our prompt learning design's proficiency in handling missing data. Additionally, different fusion strategies lead to distinct modalities integration, with the cross-fusion approach often boosting performance in specific situations, such as when dealing with missing-image cases in the Hateful Memes dataset which surpasses MPVR by about 3.53%. However, it exhibits greater sensitivity to various missing cases, particularly when text is absent. In scenarios with limited textual data, cross-fusion can inadvertently emphasize the fusion of prompts combined with modality inputs, potentially impacting multimodal representation.

4.4 Robustness Comparison.

Robustness to Different Missing Rates The performance differences in baseline models vary significantly in robustness to different missing rates. Results for various missing rates on Hateful Memes are displayed in Figure 3. Assessing robustness involves calculating the average drop rate between successive data points.

MPVR exhibits inferior performance compared to ViLT in certain cases, demonstrating the highest vulnerability with a maximum drop rate of 4.18% in the missing-text scenario and an average drop of 3.53%. Our proposed method, compared to head fusion, achieves a significant performance enhancement, with a low drop rate of only 3.05%, and average improvements of 9.76% for MPVR and 10.95% for ViLT. Our cross-fusion strategy demon-

Methods	Missing Rate ϵ	Hateful Memes (AUROC)
MuAP-w-tuning	70%	**65.09**
MuAP-w/o-single-stage		63.47
MuAP-w/o-text-step		63.65
MuAP-w/o-image-step		64.64
MuAP-w/o-KL		64.57

Table 2: Ablation study to explore how multi-step prompt tuning improves model's performance. All models using the head-fusion strategy are trained and evaluated on missing-image scenarios.

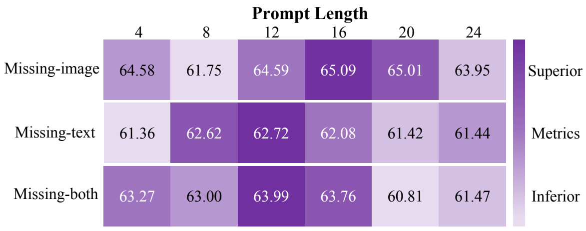


Figure 4: Ablation study on prompt length for head-fusion strategy. All models are trained and evaluated on various scenarios (e.g., missing-image) with ϵ=70%.

strates enhanced performance in most settings of the missing-image scenario, with the lowest drop rate of 2.4%. It surpasses MPVR and ViLT by an average of 8.66% and 9.85%, respectively, underscoring the effectiveness of our method in bolstering the model's resilience and performance across varying missing rate conditions.

Prompt learning enhances multimodal fusion, improving model performance. MPVR's prompting method lacks robustness, leading to overfitting and sensitivity to missing modality cases. Missing-aware ability alone is insufficient, necessitating more robust methods. Our prompt exhibits modality-specificity and achieves missing-awareness through diverse fusion techniques. Multi-step prompt tuning aligns distinct modalities via adjustments, highlighting a trade-off between model performance and robustness.

4.5 Ablation Study

Effectiveness of Multi-step Prompt Tuning One of the most innovative aspects of our approach is the multi-step prompt tuning, consisting of single-stage and alignment-stage steps. We conducted experiments to assess the impact of it. As shown in Table 2, the variation with multi-step prompt tuning achieves the best performance, while the model without any tuning performs the worst. The experiment demonstrates that without iterative tuning steps, the model fails to capture crucial modality-specific information, which is essential for effective multimodal fusion. Other variations (e.g., removing text-step, KL divergence) also show different degrees of performance decrease, indicating that this module we set up to align modalities has a significant positive effect.

Effectiveness of Prompt Length In our proposed approach, the prompt length L_P is a critical factor. For example, in the head-fusion prompting strategy, the final output length scales linearly with $(N_P L_P + L_V)$. Therefore, a judicious choice of L_P is necessary to ensure computational efficiency and prevent information disruption during the training process. We analyze the effect of prompt length in Figure 4. Consistent with intuition, model performance improves as prompt length L_P increases, peaking at values between 12 and 16. This improvement can be attributed to the additional modal information provided at shorter lengths, preventing overfitting. However, a decline in performance is observed when the length exceeds 16. This observation indicates that excessively long prompts lead to a concatenation situation where the combined length nears the original embedding length, hindering effective learning.

5 Conclusion

In this paper, we have undertaken the pioneering effort to comprehensively investigate the robustness of prompt learning models when modalities are incomplete. Our experimental findings have revealed the high sensitivity of existing prompt learning models to the absence of modalities, resulting in substantial performance degradation. Building upon these insights, we propose a Multi-step Adaptive Prompt Learning (MuAP) framework for missing-modality in the Vision-Language Model. We generate learnable modality-specific prompts and explore two prompt strategies to facilitate prompt learning in missing-modality Transformer models. To enable adaptive learning of multimodal prompts, we employ a multi-step tuning mechanism encompassing single-stage and alignment-stage tunings to perform multi-step modality alignments. This enables MuAP to acquire comprehensive knowledge from both modalities in a balanced manner. Extensive experiments conducted on benchmark datasets validate the effectiveness of MuAP.

6 Limitation

First, due to time and computational constraints, we haven't tested our techniques on LLMs and larger datasets. Second, in our choice of modalities, we've focused solely on text and visuals using ViLT. It's crucial to incorporate additional modalities such as sound. It's essential for our proposed approach to demonstrate generalizability across diverse modalities, a focus for our upcoming work. Third, we have not explored more alignment methods due to the computational limitations. Finally, despite using few parameters, the overall improvement is not substantial, but the robustness verification has significantly enhanced. Moving forward, more interpretable analysis will be carried out to comprehend the principles of the parameters' effects.

References

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, Luis Enrique Erro No, Sta Ma Tonantzintla, and Fabio A González. 2017. Gated multimodal units for information fu. *stat*, 1050:7.

Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. 2021. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7088–7097.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. 2019. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International conference on image processing (ICIP)*, pages 1440–1444. IEEE.

Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. 2019. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167:26–37.

Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. 2021. Cdnet: Complementary depth network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:3376–3390.

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.

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

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.

Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale.

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. *arXiv preprint arXiv:2104.08691*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*.

Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. 2019. Multimodal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0.

Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng. 2024. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 101:101973.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

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

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv e-prints*, pages arXiv–2001.

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.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. 2021. Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1407–1417.

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

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

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.

Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

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

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

A Implementation Details

Regarding text modality, we use the bert-base-uncased tokenizer to tokenize our input sequence. Depending on the dataset, the maximum length of text sentences is set differently. It is set to 128 for Hateful Meme, 512 for Food-101, and 1024 for MM-IMDB. For the image modality, following (Kolesnikov et al.), we extract 32×32 patches from the input image. Therefore, the input images are resized to 384×384 during the preprocessing stage.

For the missing situation, we follow (Lee et al., 2023) to keep the overall missing rate at 70%. Considering various missing scenarios, we mainly set three cases, including only the text modality (missing-text) or image modality (missing-imgae) missing $\epsilon\%$ while the other modality remains intact, and another type is both modalities (missing-both) are missing $\frac{\epsilon}{2}\%$ separately. The specific missing scenarios in training and inference experiments are shown in Table 1.

Moreover, the backbone parameters are initialized by pre-trained weights of ViLT. The length L_p of learnable prompts is set to 16 by default in both head fusion and cross fusion. We set the maximum prompt layer number to 6 (i.e. the indices of layers to pre-pend prompts start from 0 and end at 5). The base learning rate is set at 1×10^{-2} using the AdamW optimizer (Loshchilov and Hutter, 2018) and weight decay at 2×10^{-2} to remain unchanged from (Lee et al., 2023).

B Details of Various Datasets

As previously mentioned, we have three distinct datasets: MM-IMDb (Arevalo et al., 2017), UPMC Food-101 (Wang et al., 2015), and Hateful Memes (Kiela et al., 2020), each with its own objectives and evaluation metrics.

To provide a clear overview of these datasets, Figure 5 illustrates a comparison of their task objectives. MM-IMDb focuses on classifying movie genres, UMPC Food-101 is designed for food type classification, and Hateful Memes presents a formidable challenge in detecting hate speech across multiple modalities. As depicted in Figure 5, the Hateful Memes dataset poses the greatest challenge due to its extensive composition of over $10,000$ newly generated multimodal instances. The intentional selection of these instances aims to pose difficulties for single-modal classifiers in accurately labeling them. For instance, a classifier relying solely on the text “Elon Musk presents infinite energy source” may not classify it as hateful. However, when accompanied by the corresponding image of Elon Musk placing his hand on his forehead, crucial contextual information is provided to identify its hateful connotation. The tasks in MM-IMDb and UMPC Food-101 are notably less challenging due to explicit answers within the text. This is evident in the UMPC Food-101 example, where the classification result “apple pie” is directly mentioned in the text. Therefore, in our experimental setup, we primarily utilize the Hateful Memes dataset to effectively showcase the superiority of our approach compared to various baseline models.

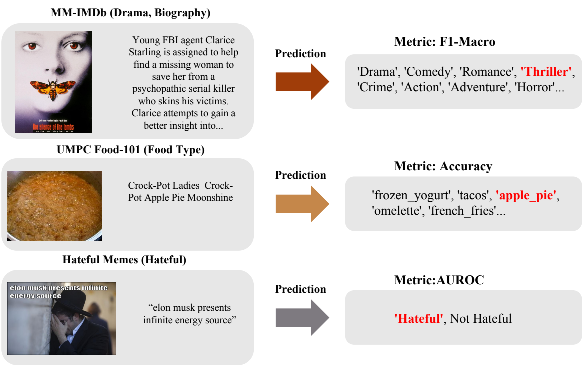


Figure 5: Detailed examples for three benchmark datasets.

C More Ablation Results

Missing-text Missing rate ϵ	ViLT	MPVR (Input-level)	MuAP (Head Fusion)	MuAP (Cross Fusion)
70%	55.56	**60.02**	57.95	58.39
50%	59.63	61.98	**68.18**	62.39
30%	65.47	67.80	**68.17**	64.85
10%	66.37	67.36	**69.79**	69.65

Table 3: Ablation study of generalization ability on Hateful Memes. All models are evaluated on missing-text cases with different missing rates ϵ.

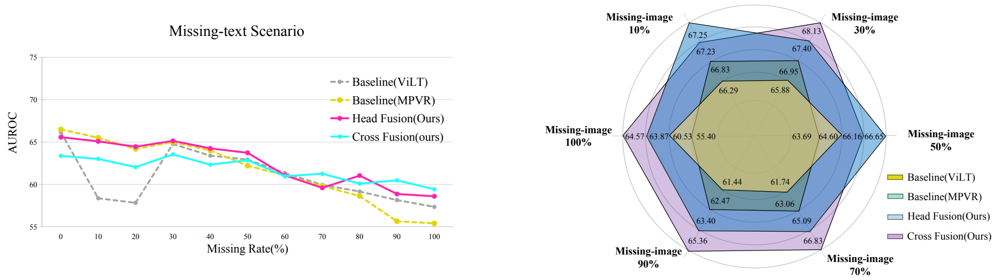
Generalization Ability Initially, we assume that real-world scenarios may involve missing modality instances due to device malfunctions or privacy concerns. However, the majority of existing datasets comprise modality-complete and meticulously annotated data. To address this inconsistency, we conducted experiments to investigate the impacts of a prompt learning model trained on complete modality datasets. In detail, all models are trained on complete modality cases and tested on scenarios with missing text at different rates. In Table 3, our findings reveal that head-fusion and cross-fusion prompting exhibit robustness to this practical situation across numerous configurations. They consistently rank among the top performers, except for ϵ values of 70%. Our head-fusion prompting strategy exhibits remarkable performance, substantially enhancing both performance and robustness in the majority of scenarios, with an average AUROC of 66.02%, which is a 1.73% improvement compared to the average performance of MPVR (64.29%). Meanwhile, the cross-fusion prompting strategy ranks second in most cases, showing a more pronounced sensitivity to specific settings compared to the head-fusion prompting strategy. According to the findings elucidated in the paper, the cross-fusion prompting strategy proves to be effective in handling incomplete multimodal data, while the head-fusion prompting strategy exhibits exceptional robustness when dealing with complete multimodal data.

Methods	Missing Rate ϵ	Hateful Memes (AUROC)
MuAP-w-tuning	70%	**66.86**
MuAP-w/o-single-stage		66.46
MuAP-w/o-text-step		64.88
MuAP-w/o-image-step		64.56
MuAP-w/o-KL		65.28

Table 4: Ablation study to explore how multi-view prompt tuning improves model’s performance. All models using the cross-fusion strategy are trained and evaluated on missing-image scenarios with missing rate ϵ=70%. Best results in **bold**.

Effectiveness Analysis in Cross-fusion Due to space limitations in the main text, our analysis focused on assessing the effectiveness of the multi-view prompt tuning module with a head-fusion strategy. To attain a more profound comprehension of our pioneering multimodal alignment method, which encompasses multiple steps for enhancing understanding, we now evaluate its effectiveness using the cross-fusion prompting strategy. As depicted in Table 4, analogous to the preceding experimental findings, the model refined with multi-view prompting exhibits exceptional performance, surpassing all comparative models, while the untuned model performs the poorest. This validation evidence underscores the significance of iterative tuning in capturing modality-specific information that is pivotal for accomplishing successful multimodal fusion.

Robustness to Different Missing Settings We conduct experiments with different missing scenarios to demonstrate our method’s robustness across



(a) Train: Missing-image 70%; Test: Missing-text (b) Train: Missing-image 70%; Test: Missing-image

Figure 6: Robustness studies conducted by varying the missing rates in different evaluation scenarios for the Hateful Memes dataset. (a) Head-fusion models are trained using the missing-image scenario with ϵ = 70%, and evaluations were performed on the opposite missing-text case. (b) All models are trained on the missing-image scenario with a 70% missing rate, and tested on consistent cases with different missing rates, representing a transition from more complete data to less complete data.

various scenarios during the training and testing process. We aim to showcase the effectiveness of our method in improving both performance and robustness.

In previous work, (Khattak et al., 2023; Liu et al., 2024) use textual modality as the main modality. So we evaluated models trained on a missing-image scenario with a 70% missing rate and diverse missing-text scenarios with varying ϵ values. Figure 6(a) shows that head-fusion consistently outperforms MPVR across scenarios, with our proposed strategies remaining robust even with increasing missing rates, achieving average AUROC values of 62.49% and 61.76% for head-fusion and cross-fusion, respectively. Our approach maintains stable performance even in highly challenging scenarios with higher missing rates, unlike MPVR, which becomes ineffective when the missing rate surpasses 80%. We attribute this improvement to our f_{missing} function, implemented using residual connections, familiarizing the model with complete and missing data scenarios, effectively facilitating information supplementation.

Figure 6(b) illustrates the robustness of our proposed method, as it consistently outperforms MPVR, particularly when tested with varying missing rates while maintaining consistent missing-image settings during training. Our multimodal prompts effectively tackle missing-awareness and modality-specificity, significantly boosting the robustness of prompt learning.

Moreover, we have analyzed the model performance in various prompt lengths with the head-fusion strategy in the main text. However, in the proposed cross-fusion prompting approach, the output length exhibits a linear increase, directly proportional to the sum of $(2N_P L_P + L_V)$. This linear

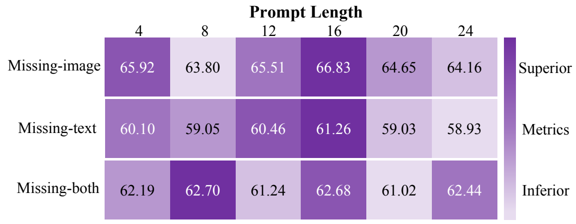


Figure 7: Ablation study on prompt length for cross-fusion strategy. All models are trained and evaluated on various scenarios (e.g., missing-image, missing-text) with ϵ=70%.

growth becomes notably more substantial as the length of the prompt escalates compared to head-fusion, which may lead to increased computational demands and potential efficiency challenges. The analysis presented in Figure 7 reveals a distinct trend compared to head-fusion. It is noteworthy that the top-3 performances in each scenario exhibit variability. Notably, when L_P ranges from 4 to 8, the performance is consistently strong, achieving the highest AUROC of 62.70% in the missing-both case. This suggests that even smaller values of L_P can yield excellent performance. However, akin to head-fusion, the optimal performance is observed when L_P approaches 16. These findings suggest that our cross-fusion method is particularly sensitive to the prompt length due to the rapid accumulation of sequence length, potentially leading to overfitting and inefficient computation.

D The Selection of Multi-view Prompt Tuning Hyperparameters

To demonstrate our selection of the involved hyperparameters, we further analyze the impact of the hyperparameters λ_t and λ_v with Hatefull Memes (Kiela et al., 2020). From Table 5, it can be

Hateful Memes

$\lambda_v \lambda_t$	0.4	0.5	0.6	0.7	AVG
0.3	65.09	64.44	65.50	66.01	65.26
0.4	64.88	63.90	64.31	64.06	64.29
0.5	64.29	65.04	64.40	63.35	64.27
0.6	64.08	64.11	64.03	63.55	63.94
AVG	64.59	64.37	64.56	64.24	64.44

Table 5: Hyperparameters selection analysis on the hyper-parameter λ for both modalities with Hateful Memes (Kiela et al., 2020). All head-fusion models are trained and tested on missing-image scenario with ϵ=70%

seen that when λ_t is 0.4, the overall performance is the best, and the same situation occurs when λ_v is 0.3. Therefore, in the remaining experiments, we maintain λ_t at 0.4 and λ_v at 0.3 to gain the maximum performance.