RobustPT: Dynamic Disentanglement Prompt Tuning in Vision-Language Models with Missing Modalities

Anonymous ACL submission

Abstract

Recently, prompt tuning has garnered considerable attention due to its success across various Vision-Language (VL) tasks. However, unimodal prompts, coupled prompts, and joint prompts in these models often lead to suboptimal performance due to differences in information density and complexity between modalities. Particularly, in scenarios with missing modalities, these prompt-based approaches tend to exacerbate 'Channel Bias'-a phonomenon where models overly rely on specific feature (such as unmissing-modal feature) channels from the base tasks, thereby undermining the model's ability to capture crucial shared knowledge applicable to new tasks and affecting its generalizability. To address this challenge, we propose **RobustPT**, a dynamic disentanglement prompt tuning model designed to enhance the robustness of VL models under modality missing conditions. RobustPT utilizes a multi-channel prompting mechanism to dynamically disentangle and align prompts. Specifically, RobustPT is divided into single-channel tuning and alignment-channel tuning, where prompts for each modality run independently in sequence to delve deeply into their intrinsic characteristics, followed by an integration through a non-strong coupling strategy to effectively balance information contributions and enhance overall performance. Extensive experiments demonstrate that our RobustPT achieve significant improvements over the current state-of-the-art across all benchmark datasets. Our codes are available at https://anonymous.4open.science/r/RobustPT-C016/.

1 Introduction

In the realm of Vision-Language (VL) pre-training, large-scale models such as CLIP (Radford et al., 2021) have achieved significant success in 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). However, existing vision-language models typically assume that all modal data are fully available during both training and testing phases. It's a highly idealized assumption that often fails to hold in practical applications. The unavailability of data due to factors such as privacy policies (Lee et al., 2023) or technological limitations (Jang et al., 2023), referred to as modality missing—has become one of the significant constraints on the performance of VL models (Su et al., 2019; Lu et al., 2019; Yu et al., 2019; Kim et al., 2021). Therefore, exploring how to maintain model performance in the face of incomplete modal data has become crucial.

Inspired by the success of prompt engineering in Natural Language Processing (NLP), prompt tuning techniques (Zhou et al., 2022b,a) have increasingly become an effective strategy for enhancing the adaptability of vision-language models to specific downstream tasks. It is achieved by optimizing prompts for specific tasks, such as using learnable unimodal prompts in (Zhou et al., 2022b; Jia et al., 2022) and multimodal coupled prompts in MaPLe (Khattak et al., 2023), where only a subset of trainable vectors are adjusted while keeping the main model weights fixed. Recently, (Lee et al., 2023) proposed a joint prompt tuning approach (MPVR) for scenarios with missing modalities, aimed at mitigating performance degradation due to modality differences in training or testing data samples. Further, to address the challenges of exponential growth in the number of prompts and robustness issues in various missing modality scenarios in MPVR, (Jang et al., 2023) proposed using modality-specific tokens as a design concept for joint prompts to simplify and enhance the model's performance across diverse modality configurations. Additionally, (Guo et al., 2024) introduced generative prompts, missing signal prompts,

and missing type prompts, effectively generating features for missing modalities and facilitating the learning of inter-modal information.

While these prompt tuning techniques have shown great potential, unimodal, coupled, and joint prompts may lead to suboptimal performance due to differences between modalities. Specifically, different modalities, such as images and text, often possess varying levels of information density and complexity, and simple coupled or joint prompt models fail to effectively balance the contributions of these modalities. Moreover, in the presence of missing modalities, model performance can deteriorate further as these systems tend to overly rely on feature channels that are particularly effective for base tasks, leading to a phenomenon known as *channel bias*. This bias hinders the model's ability to effectively capture key shared knowledge applicable to new tasks, thereby weakening the model's generalization capabilities in scenarios with missing modalities.

To address the aforementioned issues, we propose a new **Robust** method: a dynamic disentanglement **P**rompt **T**uning model designed to handle missing modalities in vision-language models, dubbed **RobustPT**. RobustPT introduces a multi-channel prompt mechanism that dynamically disentangles and aligns modalities through an iterative process, adaptively optimizing the learning of multimodal prompts. Specifically, the implementation of RobustPT is divided into two parts: single-channel learning and alignment-channel learning. In the single-channel learning phase, the prompts for each modality operate independently, unaffected by other modalities, allowing each modality to delve deeply and optimize based on its inherent characteristics. Subsequently, in the alignment-channel learning phase, the independently learned prompts are integrated using a non-strong coupling alignment strategy, maintaining the independence of each modality's prompts while achieving functional complementarity and maximizing information integration.

The main contributions of this paper are as follows:

(1) We propose **RobustPT**, a robust multi-channel dynamic disentanglement prompt tuning model that effectively addresses the issue of channel bias caused by the model's overreliance on specific modal features (such as existing modal features). By establishing separate single-channel learning and alignment-channel learning, the model can independently and adaptively generate visual and language prompts, thus better balancing the information contributions across different modalities.

(2) We employ dynamic disentanglement for prompt adjustments and a non-strong coupling strategy for alignment adjustments. By sequentially separating and aligning prompts across different learning channels, RobustPT facilitates the extraction of unique knowledge specific to foundational tasks from isolated feature spaces, while preserving shared knowledge in the original feature spaces, achieving iterative adaptability and stable generalization capabilities.

(3) We conducted extensive experiments and ablation studies on three benchmark datasets to validate the effectiveness of the RobustPT model. These experiments demonstrate that our model achieves significant performance improvements over existing state-of-the-art technologies across all benchmark datasets, showcasing its superior capabilities in handling missing modality issues.

2 Related work

2.1 Prompt Tuning 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 tuning 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 tuning 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.

2.2 Prompt Tuning for Missing-Modality Vision-Language Task

While previous studies have demonstrated the effectiveness of prompt tuning in VL tasks, there is limited literature on prompt tuning methods for scenarios involving missing modalities. MPVR (Lee et al., 2023) introduces a groundbreaking approach by incorporating missing modality-aware prompts in ViLT to address scenarios with missing modalities. 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 (a scenario with C modalities necessitates $2^C - 1$ prompts), posing a significant challenge to the model's scalability.

To tackle this challenge, (Jang et al., 2023) proposed modality-specific tokens as a design strategy for joint prompts, simplifying the model's architecture and enhancing performance across various modality configurations. Additionally, (Guo et al., 2024) introduced generative prompts, missing signal prompts, and missing type prompts, effectively generating features for absent modalities and promoting the learning of cross-modal information. However, these prompt models struggle to effectively balance the contributions of different modalities, often resulting in suboptimal performance.

3 Method

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

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 tuning with missing modalities $\mathcal{R}$.

3.2 Overall Framework

We propose RobustPT, which is designed to address the challenges posed by missing modalities in Vision-Language (VL) models. Following (Lee et al., 2023; Jang et al., 2023), we focus on the VL model with prompt tuning and adopt Vision-and-Language Transformer (ViLT) (Kim et al., 2021) as the backbone, which is pre-trained on large-scale VL datasets and remains untrainable in downstream tasks. As illustrated in Figure 1, RobustPT mainly comprises three modules: Multimodal Prompt Generator, Prompt Strategy Design, and Multi-channel Prompt Tuning. Specifically, we first generate learnable specific prompts for each modality to achieve completeness tuning in prompting, deviating from previous methods (Khattak et al., 2023) 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-channel tuning strategy for dynamic text/image-channel and alignment-channel prompt tuning through modality disentanglement and alignment, allowing RobustPT to gain knowledge from both modalities while maintaining the independence of each modality's prompts.

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. Specifi-

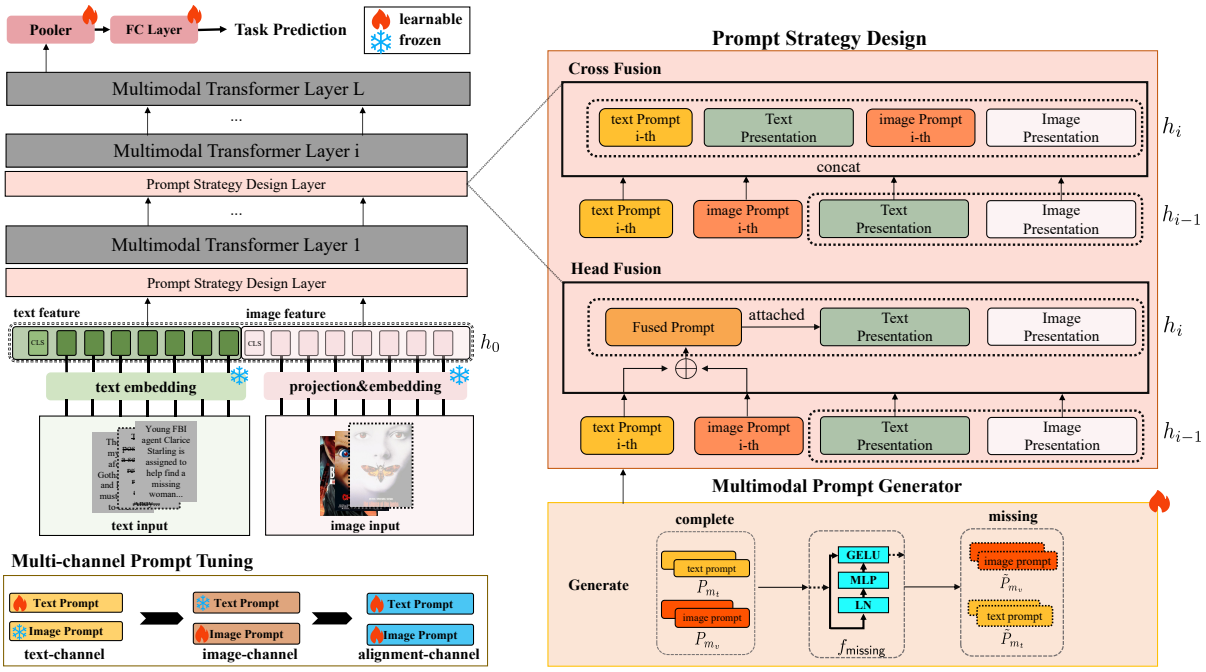


Figure 1: The overview of our RobustPT framework. The Multimodal Prompt Generator initially generates complete-type prompts, P_{m_t} and P_{m_v}, tailored to the specific modality case (e.g., textual or visual modalities in Vision-Language tasks). Next, it employs f_{missing} to create missing-type prompts $\tilde{P}_{m_t}$ 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-channel Prompt Tuning to synchronize distinct characteristics of different modality prompts effectively.

cally:

$$h^0 = [t + t^{modal}; v + v^{modal}] \in \mathbb{R}^{\mathbb{L}_{\mathbb{V}} \times \mathrm{d}}, \quad (1)$$

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

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

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 final output context sequence h^L and classification token t_{cls} are utilized for downstream tasks.

3.4 Multimodal Prompt Generator

One main challenge in addressing missing modality learning with prompt tuning 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 $P_{m_v} \in \mathbb{R}^{L_p \times d}$ 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 1: 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:

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

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:

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

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-channel Prompt Tuning

In this section, we introduce our proposed multi-channel prompt tuning technique designed to adaptively learn multimodal prompts through multi-channel sequential modality alignments. Specifically, We utilize prompt tuning of the pre-trained Transformer encoder for efficient parameter learning across multiple stages, enabling the extraction of modality-specific information and capturing correlations between visual and textual modalities.

Single-channel 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 1, we iteratively train the learnable prompts in a channel-wise manner. Initially, we optimize the textual prompts while keeping the visual prompts frozen, called text-channel. Subsequently, we switch to optimizing the visual prompts while fixing the textual prompts, called image-channel.

This exclusive updating process enables the prompt tuning to capture modality-individual attributes respectively.

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

$$\mathcal{L}_{cls} = -\sum_{k=1}^{N} y_k \log(\hat{y}_k), \tag{10}$$

$$\mathcal{L}_{kl}(P_t, P_v) = \sum_{i=1}^{N_p} \sum_{k=1}^{N} P_t^i(k) \log\left(P_t^i(k)/P_v^i(k)\right), \tag{11}$$

where P_{ti} and P_{vi} which are employed across N_P layers to calculate the KL divergence for enhancing modal-correlative information. Assuming the dataset consists of N samples, each belonging to one of T classes. $\hat{y}_{it}$ and y_{it} denote the predict probability and actual probability of sample i belonging to class t, with cross-entropy loss as the main optimization criterion.

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

$$\textbf{Text-channel} : \mathcal{L}_{total}^{t} = \mathcal{L}_{cls} + \lambda_t \mathcal{L}_{kl}(P_{m_t}, P_{m_v}), \tag{12}$$

$$\textbf{Image-channel} : \mathcal{L}_{total}^{v} = \mathcal{L}_{cls} + \lambda_v \mathcal{L}_{kl}(P_{m_t}, P_{m_v}). \tag{13}$$

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 channel and 0.3 for the image prompt training channel, respectively. In the process of single-channel prompt tuning, the two prompts undergo simultaneous updates through several alignment channels, with the experimental setup setting the number of channels to 3.

Alignment-channel 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 channel, 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:

$$\textbf{Alignment-channel} : \mathcal{L}_{total} = \mathcal{L}_{cls}. \tag{14}$$

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), UPMC Food-101 (Wang et al., 2015) and Hateful Memes (Kiela et al., 2020). Due to page limitations, more details can be found in Appendix B.

Metrics Given the distinct classification tasks addressed by these datasets, we employ appropriate metrics tailored to each one. Specifically, for MM-IMDb, we utilize F1-Macro to measure 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 Wcompare it with the SOTA models (Appendix A):

- ***Finetuned VILT (Kim et al., 2021)***: the original one without additional parameters in ViLT (i.e. only training the pooler layer and task-specific classifier).
- ***Visual BERT (Li et al., 2019)***: modified by only focusing on pooler and classifier training.
- ***MPVR (Lee et al., 2023)***: derived from the pre-trained VILT backbone, this model integrates **missing-aware prompts** into its multimodal transformer design.
- ***MSPs (Jang et al., 2023)***: using modality-specific tokens to combine prompts and reduce prompt parameters is a powerful competitive baseline.

4.3 Main Results

Basic Performance. Table 1 shows our new prompt tuning 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

Datasets	Missing rate ϵ	Training		Testing		ViLT	Visual BERT	MPVR (Input-level)	MPVR (Attention-level)	MSPs (Jang et al., 2023)	RobustPT (Head Fusion)	RobustPT (Cross Fusion)
		Image	Text	Image	Text							
MM-IMDb (F1-Macro)	70%	30%	100%	30%	100%	37.61	38.63	46.30	44.74	47.45	**48.21**(+22.0%)	46.73(+5.0%)
		65%	65%	65%	65%	36.30	37.23	42.41	41.56	42.03	42.57(+3.0%)	**43.92**(+25%)
		100%	30%	100%	30%	34.71	36.41	39.19	38.13	38.34	**41.37**(+48.7%)	39.88(+15.4%)
Food101 (Accuracy)	70%	30%	100%	30%	100%	76.93	77.41	86.09	85.89	86.34	**86.90**(+8.8%)	86.59(+5.5%)
		65%	65%	65%	65%	69.03	71.06	77.49	77.55	78.89	77.87(+4.5%)	**78.95**(+17.3%)
		100%	30%	100%	30%	66.29	67.78	73.85	72.47	73.77	**74.61**(+10.1%)	74.60(+9.9%)
Hateful Memes (AUROC)	70%	30%	100%	30%	100%	61.74	61.98	62.34	63.30	-	65.09(+458.3%)	**66.83**(+748.3%)
		65%	65%	65%	65%	62.83	63.05	63.53	62.56	-	**64.76**(+175.5%)	62.68(-121.4%)
		100%	30%	100%	30%	60.83	60.89	61.01	61.77	-	**62.08**(+594.4%)	61.26(+138.9%)

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 $\epsilon\% = 70\%$. The outcomes were analyzed under diverse missing-modality cases, with the best results highlighted in **bold** for clarity. The percentages in green and red indicate the increment/decrement of our method: (RobustPT - MPVR(input-level)) / (MPVR(input-level) - ViLT).

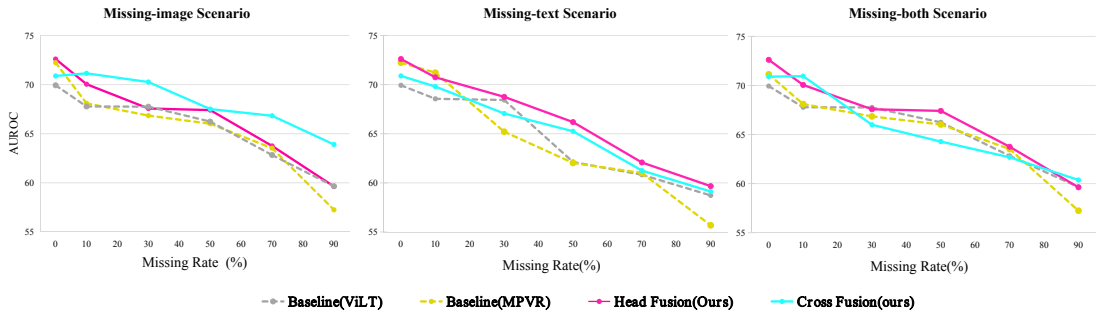


Figure 2: Comparison of baselines on the Hateful Memes dataset with different missing rates across various missing-modality scenarios. Each point in the picture represents training and testing with the same $\epsilon\%$ missing rate.

dataset, showing a 1.94% average improvement. In MM-IMDB, our best method shows an average improvement of 1.89% over MSPs. This highlights our prompt tuning 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 baselines vary significantly in robustness to different missing rates. Figure 2 shows results for different missing rates on Hateful Memes. Robustness is evaluated by 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 missing-text scenario and an average drop of 3.53%. Our proposed 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 demonstrates 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. From a statistical perspective, under different missing rates, the head-fusion and cross-fusion of RobustPT achieved averages of 66.80 and 66.62, while MPVR only reached 65.17. Additionally, the range for cross-fusion was 9.86, with a standard deviation of 3.61, while MPVR had a range of 14.11 and a standard deviation of 4.63, underscoring the effectiveness of our method in bolstering the model's resilience and performance across varying missing conditions.

Prompt tuning enhances multimodal fusion, improving model performance. MPVR's prompting method lacks robustness, leading to overfit-

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

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

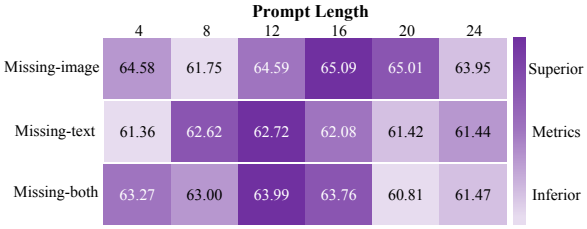


Figure 3: Ablation study on prompt length for head-fusion strategy. All models are trained and evaluated on various scenarios (e.g., missing-image) with ϵ=70%.

ting 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-channel prompt tuning aligns distinct modalities via adjustments, highlighting a trade-off between model performance and robustness.

4.5 Ablation Study

Effectiveness of Multi-channel Prompt Tuning One of the most innovative aspects of our approach is the multi-channel prompt tuning, consisting of single-channel and alignment-channel channels. We conducted experiments to assess the impact of it. As shown in Table 2, the variation with multi-channel prompt tuning achieves the best performance, while the model without any tuning performs the worst. The experiment demonstrates that without iterative tuning channels, the model fails to capture crucial modality-individual information, which is essential for effective multimodal fusion. Other variations (e.g., removing text-channel, 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 3. 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. 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 explore the robustness of prompt tuning techniques in Vision-Language (VL) models under incomplete modalities.Although traditional prompt tuning techniques have been successful in various VL tasks, common approaches such as single-modality, coupled, and joint prompt tuning often underperform due to discrepancies in information density and complexity when modalities are missing. To address these challenges, we propose RobustPT, a dynamic disentanglement prompt tuning model designed to enhance the robustness of VL models facing modality missing. RobustPT employs dynamic disentanglement and non-strong coupling alignment strategies to iteratively perform separated prompt tuning and aligned prompt tuning, facilitating the extraction of unique knowledge while maintaining crucial shared insights for adaptive and stable generalization. Our extensive experiments across multiple benchmark datasets validate the effectiveness of RobustPT, which shows significant performance improvements over existing missing-modality methods.

6 Limitation

First, in our choice of modalities, we've focused solely on text and visuals using ViLT. Our method's effectiveness has been proven through extensive experiments. However, 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. Second, we have not yet explored additional alignment methods to assess their impact on multimodal information, leaving this for future

work.

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.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. *Preprint*, arXiv:2407.05374.

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.

Jaehyuk Jang, Yooseung Wang, and Changick Kim. 2023. Towards robust multimodal prompting with missing modalities. *Preprint*, arXiv:2312.15890.

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

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.

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*.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

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.

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.

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. Notably, different devices may lead to different results, and we also need to determine whether the experimental effects of MPVR are usable. This is because the results in Table 1 of MPVR appear to have an outlier, i.e. the missing-both scenario is much higher than the missing-image scenario in hateful memes, while in other datasets, the missing-image is always the highest with a 70% missing rate. We reproduced the results based on the official code of MPVR [1] and the same experimental parameters . Our results are comparable to the original paper, with some results being better and some being worse. As the outlier problem mentioned above, we have never successfully reproduced it after multiple attempts, so a baseline based on reproduction is more trustworthy to prove the effectiveness of our method.

Moreover, we use Pytorch (Paszke et al., 2019) for implementation. The backbone parameters are initialized by pre-trained weights of ViLT (87.4M). 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.

- ***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.

To provide a clear overview of these datasets, Figure 4 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

[1] https://github.com/YiLunLee/missing_aware_prompts

across multiple modalities. As depicted in Figure 4, 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.

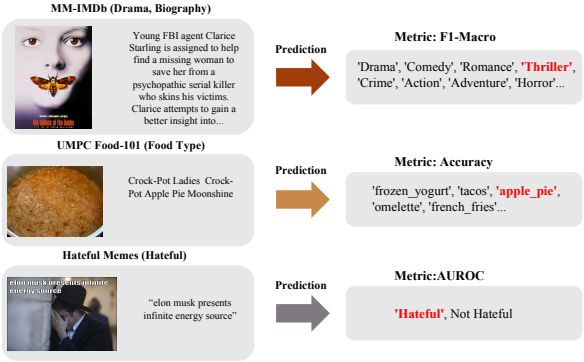


Figure 4: Detailed examples for three benchmark datasets.

C More Ablation Results

Missing-text Missing rate ϵ	ViLT	RobustPT (Input-level)	RobustPT (Head Fusion)	RobustPT (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 tuning 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)
RobustPT-w-tuning	70%	**66.86**
RobustPT-w/o-single-stage		66.46
RobustPT-w/o-text-channel		64.88
RobustPT-w/o-image-channel		64.56
RobustPT-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 channels 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 multi-modal fusion.

Robustness to Different Missing Settings We conduct experiments with different missing scenarios to demonstrate our method's robustness across 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 5(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 5(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 tuning.

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 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 6 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.

Training	Testing	MPVR	RobustPT
missing-image 100%	missing-image 100%	**62.01**	61.94
missing-text 100%	missing-text 100%	59.00	**60.40**
missing-image 70%	missing-text 100%	55.40	**58.60**
missing-image 70%	missing-image 100%	66.89	**67.25**
Average		60.83	**62.05**

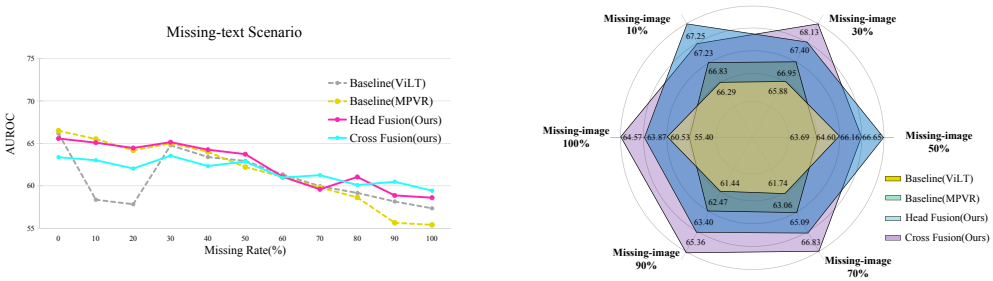
Table 5: Performance comparison based on single modality

Performance on single modality We conducted several experiments by varying the missing rates up to 100% in single-modality testing across different evaluation scenarios. For example, we had a missing-image at the training stage of 70% with and a missing-image of 100% at the testing stage. However, these experiments do not fully address how models can learn characteristics from a completely absent modality. Hence, we conducted new experiments with two single-modality scenarios: missing-image 100% and missing-text 100%. Even under these more challenging conditions, our RobustPT still achieved better results than MPVR. The results for hateful memes are shown in Table 5.

Hateful Memes

λ_v \ λ_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 6: 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%



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

Figure 5: 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.

	Prompt Length					
	4	8	12	16	20	24
Missing-image	65.92	63.80	65.51	66.83	64.65	64.16
Missing-text	60.10	59.05	60.46	61.26	59.03	58.93
Missing-both	62.19	62.70	61.24	62.68	61.02	62.44

Superior
Metrics
Inferior

Figure 6: 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%.

D The Selection of Multi-channel 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 6, it can be 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.