Retrieval-Augmented Dynamic Prompt Tuning for Incomplete Multimodal Learning

Jian Lang[1*], **Zhangtao Cheng**[1*], **Ting Zhong**[1,2], **Fan Zhou**[1,2†]

[1]University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[2]Kash Institute of Electronics and Information Industry, Kashgar, Xinjiang, China
jian_lang@std.uestc.edu.cn, zhangtao.cheng@outlook.com, zhongting@uestc.edu.cn, fan.zhou@uestc.edu.cn

Abstract

Multimodal learning with incomplete modality is practical and challenging. Recently, researchers have focused on enhancing the robustness of pre-trained MultiModal Transformers (MMTs) under missing modality conditions by applying learnable prompts. However, these prompt-based methods face several limitations: (1) incomplete modalities provide restricted modal cues for task-specific inference, (2) dummy imputation for missing content causes information loss and introduces noise, and (3) static prompts are instance-agnostic, offering limited knowledge for instances with various missing conditions. To address these issues, we propose RAGPT, a novel **R**etrieval-**A**u**G**mented dynamic **P**rompt **T**uning framework. RAGPT comprises three modules: (I) the multi-channel retriever, which identifies similar instances through a within-modality retrieval strategy, (II) the missing modality generator, which recovers missing information using retrieved contexts, and (III) the context-aware prompter, which captures contextual knowledge from relevant instances and generates dynamic prompts to largely enhance the MMT's robustness. Extensive experiments conducted on three real-world datasets show that RAGPT consistently outperforms all competitive baselines in handling incomplete modality problems.

Introduction

Multimodal learning has emerged as a critical paradigm in both research and industry, demonstrating broad application potential in areas such as healthcare assistance (Ghosh et al. 2024) and malicious content detection (Kiela et al. 2020). However, most successful methods typically assume that the completeness of all modalities is essential during both training and inference phases. In reality, factors such as malfunctioning sensors and privacy concerns often make it infeasible to collect complete modalities (Ma et al. 2021). As a result, the challenge of incomplete modalities significantly impacts the reliability, accuracy, and safety of multimodal models in practical applications (Woo et al. 2023; Cheng et al. 2024a).

To address this challenge, researchers have developed various robust multimodal methods that are broadly categorized into three groups: *(1) Joint learning methods* (Wang et al. 2023; Yao et al. 2024), *(2) Cross-modal generation methods* (Ma et al. 2021; Woo et al. 2023), and *(3) Prompt-based methods* (Lee et al. 2023; Jang, Wang, and Kim 2024). For joint learning methods, they heavily rely on the selection of similarity measures and require filling missing-modality inputs with masking values, resulting in the loss of critical information and the introduction of noise into the models (Wang et al. 2024). Cross-modal generation methods inevitably face modality heterogeneity issues and incur limited reconstruction quality. Recently, prompt-based methods have gained significant attention due to the rise of powerful pre-trained MultiModal Transformers (MMTs). These methods leverage prompt-tuning techniques to effectively transfer the capabilities of MMTs pre-trained on complete multimodal datasets to tasks involving missing modalities, achieving remarkable performance and making them a dominant trend in incomplete multimodal learning.

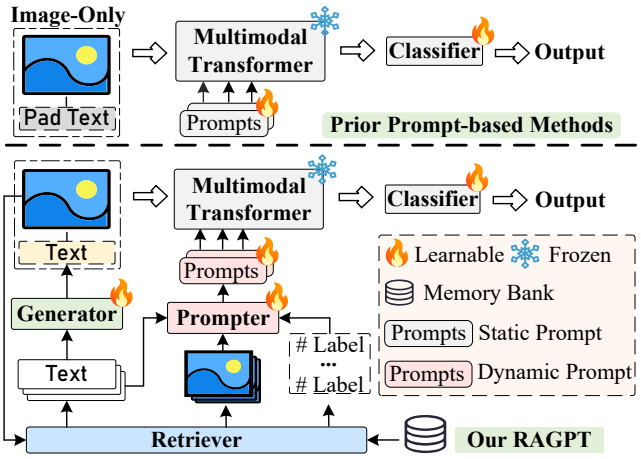


Figure 1: Prior prompt-based methods *vs.* our RAGPT in tackling incomplete multimodal learning.

However, for incomplete modalities, prompt-based methods typically use the available modalities as the only cue to fulfill task-specific objectives through prompt learning (see Fig. 1). Despite their progress, these methods often struggle in severe missing-modality scenarios due to several unresolved issues inherent in their design: **(1)** Remaining modalities typically provide restricted modal informa-

*These authors contributed equally.
†Corresponding author.

tion, which fails to effectively address specific tasks when the missing modality contains crucial modal cues. **(2)** Modal incomplete inputs are often filled with dummy values (e.g., empty strings/pixels for texts/images), which may introduce noise, leading to degraded performance (Ma et al. 2022). **(3)** The prompt tokens are shared across any inputs and therefore are instance-agnostic. Thus, this static prompt-tuning is not well-suited for real multimodal instances, as instances with different types of missing modalities belong to distinct distributions. Additionally, static prompts typically provide limited knowledge for both missing- and full-modality instances. Therefore, these observations motivate us to design a universal prompt-tuning strategy to enhance the pre-trained MMT's robustness for incomplete modalities.

To address these issues, we draw inspiration from the human ability to learn through observation, which involves mastering skills by observing relevant subjects rather than attempting to memorize every subject (Hodges et al. 2007). As shown in Fig. 1, we leverage this cognitive principle to address the challenge of missing modalities. Our core idea is to retrieve relevant multimodal contents and utilize them as prompts to enhance the robustness of pre-trained MMT in both missing- and full-modality scenarios. Intuitively, for instances with missing modalities, appending multimodal content from similar instances can provide contextual knowledge relevant to the missing modality and improve task-specific predictions.

To this end, we propose **RAGPT**, a novel **R**etrieval-**A**u**G**mented dynamic **P**rompt **T**uning framework to adaptively enhance the robustness of pre-trained MMT in both missing- and full-modality scenarios. Fundamentally, we reformulate incomplete modality learning in a principled retrieve-and-prompt manner and maintain a model-agnostic design that facilitates seamless integration with various prompt-based models. RAGPT includes three modules: multi-channel retriever, missing modality generator, and context-aware prompter. During retrieval, we propose a universal multi-channel retrieval strategy that disentangles multimodal representations into unimodal components, facilitating the retrieval of similar samples based on within-modality similarities for missing- and full-modality scenarios.

Next, the missing modality generator comprises a learnable filter to approximate the missing information. Beyond traditional reconstruction techniques, which suffer from modality gaps during the cross-modal generation, this generator realizes *intra-modal reconstruction* by leveraging information from the retrieved samples that belong to the same modality as the missing one to recover the missing content. Moreover, this design enriches the missing-modality representation, ensuring alignment with the complete-modality input format of pre-trained MMTs during the pre-training phase. Finally, the context-aware prompter identifies the semantic correlations between the target and retrieved instances, producing dynamic multimodal prompts tailored to different inputs. These prompts facilitate the adaptive refinement of modality features in both missing- and full-modality scenarios, thereby enhancing the robustness of the pre-trained models. We insert these modules into the pre-trained MMTs to achieve a more accurate representation for both missing- and full-modality data. Following are our main contributions:

- To our best knowledge, this is the first retrieval-augmented paradigm for incomplete modalities. We reveal that prior prompt-based methods suffer from issues related to dummy padding and static prompts, which drastically degrade performance in severe missing-modality cases.
- To address these issues, we propose RAGPT, pioneering a retrieval-augmented dynamic prompt-tuning framework that bridges target and relevant instances, recovers missing modality, and generates dynamic prompts to enhance the MMT's robustness in diverse missing-modality situations.
- We conduct extensive experiments on three real-world datasets to evaluate RAGPT in comparison with 9 competitive baselines and the results confirm RAGPT's effectiveness in addressing missing-modality issues. The code of our work and prompt-based baselines is available at https://github.com/Jian-Lang/RAGPT.

Related Work

Incomplete Multimodal Learning. Researchers have developed various methods for incomplete multimodal learning, which can be divided into three groups: *(1) Joint learning methods* (Zhao, Li, and Jin 2021; Wang et al. 2023; Yao et al. 2024) focus on distilling complex correlations from complete modalities to tackle missing-modality data. However, these methods require filling modality-incomplete inputs with masking values, which may cause unexpected behavior and introduce additional noise. *(2) Cross-modal generation methods* (Lee et al. 2019; Yuan et al. 2021) primarily reconstruct the missing content by using remaining modalities. Researchers (Ma et al. 2021; Woo et al. 2023) directly employ VAE to generate the missing-modality based only on available modalities. Consequently, these methods inevitably face modality heterogeneity problems. *(3) Prompt-based methods* (Lee et al. 2023; Jang, Wang, and Kim 2024) introduces learnable prompts to help pre-trained MMTs address incomplete modalities.

However, prompt-based methods are constrained by the dummy imputation and static prompting strategy, resulting in performance bottlenecks. In contrast, our RAGPT captures contextual knowledge from retrieved instances to recover the missing content and generate dynamic prompts to enhance the MMT's robustness for missing modalities.

Prompt Learning. Prompt learning (Liu et al. 2023) utilizes a small number of learnable prompt parameters added to the input of pre-trained transformers, facilitating adjustments to the pre-trained models for alignment with downstream tasks. It has been successfully applied to various domains, such as visual identity (Khattak et al. 2023; Lee et al. 2023) and social network analysis (Zhou et al. 2021; Xu et al. 2021; Zhong et al. 2024; Cheng et al. 2024b, 2023). Following the success of prompt learning in NLP tasks (Li and Liang 2021), recent works have attempted to explore its application in multimodal learning (Zhou et al. 2022a). For instance, MaPLe (Khattak et al. 2023) introduces a soft prompt appended to the hidden representations of MMTs, resulting in significant improvements in few-shot image recognition. For incomplete multimodal learning, MAPs (Lee et al.

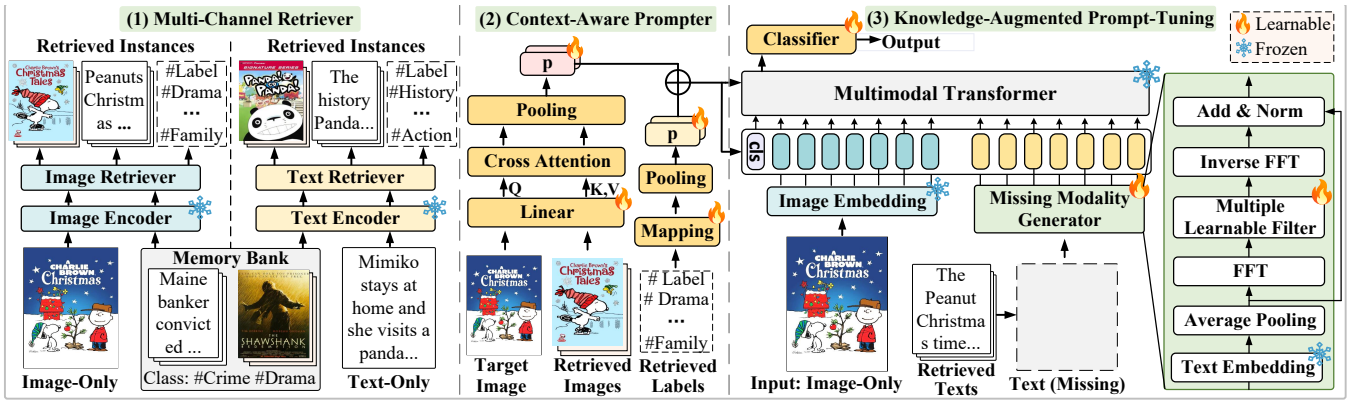


Figure 2: Overall framework of RAGPT. (1) The multi-channel retriever identifies similar instances through a within-modality retrieval strategy; (2) The context-aware prompter captures contextual knowledge from relevant instances and generates dynamic prompts; (3) The knowledge-augmented prompt-tuning process first recovers the missing content by using a missing-modality generator and then performs dynamic prompt-tuning on the pre-trained MMT for final prediction.

2023) and MSPs (Jang, Wang, and Kim 2024) design various prompts to fine-tune pre-trained MMTs, enabling them to adapt effectively to missing-modality scenarios. However, these prompts are instance-agnostic and provide limited information for both missing- and full-modality data. In contrast, the context-aware prompter in RAGPT captures rich contextual knowledge from relevant instances, alleviating the drawbacks associated with instance-agnostic prompts.

Methodology

Problem Definition: In this paper, we consider a multimodal dataset incorporating two modalities. Formally, we define $D = \{D^f, D^m\}$ to represent the multimodal dataset. Here, $D^f = \{(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i)\}_{i=1}^{N^f}$ represents the modality-complete subset, where y_i is the class label of i-th instance. $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_i^2$ denote two modalities (e.g., texts and images). N^f is the total number of instances in the subset D^f. Conversely, $D^m = \{(\boldsymbol{x}_i^1, \_\_, y_i) \vee (\_\_, \boldsymbol{x}_i^2, y_i)\}_{i=1}^{N^m}$ is a modality-incomplete subset, where "$\_\_$" indicates a missing modality and N^m is the number of missing-modality data in D^m. The objective of the task is to enhance model robustness in cases where modalities are missing during both training and testing phases.

Fig. 2 presents the key components and their relationships in RAGPT. The following sections delve into the specifics of each component and their respective implementations.

Multi-Channel Retriever

In this section, we design a unified multi-channel retriever to identify similar modal content for queries within their respective modalities by using within-modality similarities.

Memory Construction To store high-quality semantic information as prior knowledge, we define the memory $\mathcal{B}$, which encodes multimodal instances using a collection of $(\text{image}, \text{text}, \text{label})$ triples.

Multi-Channel Retrieval To adapt diverse missing- and full-modality scenarios, we develop a Multi-Channel Retriever (MCR) that effectively retrieves relevant instances through a unified retrieval architecture. Specifically, for the text-missing channel, the MCR employs the image representation as a query to identify top-K similar images and incorporates the associated texts to create multimodal instances. For complete modalities, the MCR utilizes both the image and text to search relevant texts and images, respectively, thereby creating multimodal instances.

Specifically, in the text-level branch, the MCR first tokenizes the text $\boldsymbol{x}_i^1$ in the target instance $\mathcal{T}_i$ into n word tokens and then projects them into word embedding $\mathcal{W}_i \in \mathbb{R}^{n \times d_t}$, where d_t is the dimension of word embedding. Next, the embedding $\mathcal{W}_i$ is fed into a pre-trained textual encoder (e.g., CLIP textual encoder (Radford et al. 2021)) $\Psi_t(\cdot)$ to obtain text representation, represented as $\mathbf{E}_i^t = \Psi_t(\mathcal{W}_i) \in \mathbb{R}^{d_t}$. Subsequently, the MCR utilizes the text query $\mathbf{E}_i^t$ to calculate similarity scores with the text representation $\mathbf{E}_r^t$ from the memory $\mathcal{B}$, enabling the identification of the top-K textually similar instances $\mathcal{C}_i^{\mathcal{R}}$:

$$\mathcal{C}_i^{\mathcal{R}} = \underset{r \in \mathcal{B}}{\text{Top-}K}\left(\frac{\mathbf{E}_i^{t\top}\mathbf{E}_r^t}{\|\mathbf{E}_i^t\| \cdot \|\mathbf{E}_r^t\|}\right). \tag{1}$$

For the vision content, the MCR first divides the image $\boldsymbol{x}_i^2$ into m non-overlapping patches and then projects them into a sequence of patch tokens $\mathcal{V}_i \in \mathbb{R}^{m \times d_v}$. Next, these tokens $\mathcal{V}_i$ are input into a pre-trained vision encoder (e.g., CLIP vision encoder (Radford et al. 2021)) $\Psi_v(\cdot)$ to obtain vision query $\mathbf{E}_i^v \in \mathbb{R}^{d_v}$. Finally, the retrieval process for searching top-K vision content is the same as defined in Eq. 1. After retrieval, the top-K instances $\mathcal{C}_i^{\mathcal{R}} = \{c_i^{r_1}, \cdots, c_i^{r_K}\}$ can be readily obtained. Each retrieved instance $c_i^{r_k}$ contains the $(\text{image}, \text{text}, \text{label})$ triplet. The retrieved top-K instances provide auxiliary context, guiding the recovery of missing content in the target instance and improving task-specific predictions.

Context-Aware Prompter

To explicitly capture expressive contextual information and enhance robustness of pre-trained MMTs against missing-

modality issues, we design a Context-Aware Prompter (CAP) that constructs text-, vision-, and label-level dynamic prompts from the retrieved instances $\mathcal{C}_i^{\mathcal{R}}$. For text-level prompts, the CAP fuses the reference textual features in $\mathcal{C}_i^{\mathcal{R}}$ and aligns textual embedding in $\mathcal{T}_i$ through a simple network. Specifically, the CAP first tokenizes and projects the texts $\boldsymbol{x}_i^1$ and $\{\boldsymbol{x}_i^{1,r_k}\}_{k=1}^K$ into word embeddings $\mathcal{W}_i \in \mathbb{R}^{n\times d_t}$ and $\mathcal{W}_i^{\mathcal{R}} = \{\mathcal{W}_i^{r_k}\}_{k=1}^K \in \mathbb{R}^{K\times n\times d_t}$. Subsequently, the word embedding $\mathcal{W}_i$ is used as the query to interact with the retrieved text features $\{\mathcal{W}_i^{r_k}\}_{k=1}^K$ via a cross-attention block to facilitate comprehension of context, thereby generating the text-level comprehensive representation $\tilde{\mathbf{P}}_i^t \in \mathbb{R}^{n\times d_t}$:

$$\tilde{\mathbf{P}}_i^t = \text{Att}\left(f_t^Q(\mathcal{W}_i), f_t^K(\mathcal{W}_i^{\mathcal{R}}), f_t^V(\mathcal{W}_i^{\mathcal{R}})\right), \tag{2}$$

$$\text{Att}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \tag{3}$$

where $f_t^Q(.), f_t^K(.), f_t^V(.)$ denote the query, key, and value projection functions, respectively. For vision-level prompts, the CAP uses the same process to interact the vision patch tokens $\mathcal{V}_i \in \mathbb{R}^{m\times d_v}$ with the retrieved patch tokens $\mathcal{V}_i^{\mathcal{R}} \in \mathbb{R}^{K\times m\times d_v}$ to obtain the vision-level representation $\tilde{\mathbf{P}}_i^v \in \mathbb{R}^{m\times d_v}$. Then, the CAP employs an adaptive pooling strategy to obtain the final context-aware prompts $\mathbf{P}_i^t \in \mathbb{R}^{l\times d_t}$ and $\mathbf{P}_i^v \in \mathbb{R}^{l\times d_v}$, where l is the prompt length. For label-level prompts, the CAP yields a label embedding matrix $\tilde{\mathbf{P}}_i^l \in \mathbb{R}^{C\times d}$ to encode C class labels, where d is an adjustable dimension. Given retrieved labels, the CAP performs a look-up operation on embedding matrix $\tilde{\mathbf{P}}_i^l$ and obtains each label embedding. Next, the CAP averages K label embeddings and generates label-level prompts $\mathbf{P}_i^l \in \mathbb{R}^d$.

Knowledge-Augmented Prompt-Tuning

In this process, we first utilize the retrieved modal information to approximate the missing content through a missing modality generator. Next, we perform dynamic prompt-tuning on the pre-trained MMT (e.g., ViLT (Kim, Son, and Kim 2021)) to enhance task-specific inference.

Missing Modality Generator Existing reconstruction methods (Ma et al. 2021) address missing-modality issues by recovering missing content through available modalities. However, these methods often overlook the modal heterogeneity issue and rely on complex generative structures. Based on these observations, we propose a Missing Modality Generator (MMG) that recovers the missing modality through an "intra-modal reconstruction". The MMG leverages retrieved content of the same modality as the missing one and incorporates a learnable filter layer to effectively approximate the missing modality in a simpler but effective manner. Specifically, given the text-missing instance $\mathcal{T}_i$, the MMG employs a non-parametric strategy to average all text embeddings $\mathcal{W}_i^{\mathcal{R}} = \{\mathcal{W}_i^{r_k}\}_{k=1}^K$ from retrieved instances $\mathcal{C}_i^{\mathcal{R}}$, thereby obtaining textual representation $\bar{\mathcal{W}}_i \in \mathbb{R}^{n\times d_t}$ to approximate the missing modality.

Considering potential noise in comprehensive textual representation $\bar{\mathcal{W}}_i$, the MMG introduces a simple learnable filter block (i.e., MLP-based filter (Zhou et al. 2022b)) to efficiently refine textual features $\bar{\mathcal{W}}_i$ by removing noise. Specifically, the MMG employs the Fast Fourier Transform (FFT) along the textual dimension. This operation transforms the text context representation $\bar{\mathcal{W}}_i$ into the frequency domain:

$$\mathbf{Z}_i = \mathcal{F}(\bar{\mathcal{W}}_i) \in \mathbb{C}^{n\times d_t}, \tag{4}$$

where $\mathcal{F}(\cdot)$ denotes the one-dimensional FFT, and $\mathbf{Z}_i$ is the spectrum of $\bar{\mathcal{W}}_i$. The MMG then modulates the spectrum by element-wise multiplication with a learnable filter $\mathbf{W} \in \mathbb{C}^{n\times d_t}$:

$$\tilde{\mathbf{Z}}_i = \mathbf{W} \odot \mathbf{Z}_i, \tag{5}$$

where $\odot$ denotes the element-wise multiplication. Finally, the MMG utilizes the inverse FFT operation to the modulated spectrum $\tilde{\mathbf{Z}}_i$ back into the time domain:

$$\tilde{\mathcal{W}}_i = \mathcal{F}^{-1}(\tilde{\mathbf{Z}}_i) \in \mathbb{R}^{n\times d_t}, \tag{6}$$

where $\mathcal{F}^{-1}(\cdot)$ is the inverse one-dimensional FFT, converting the complex tensor back to a real-valued tensor. To further stabilize training and enhance the embedding, the MMG incorporates a skip connection, layer normalization, and dropout:

$$\hat{\mathcal{W}}_i = \text{LayerNorm}(\tilde{\mathcal{W}}_i + \text{Dropout}(\tilde{\mathcal{W}}_i)). \tag{7}$$

Finally, the recovered representation $\hat{\mathcal{W}}_i$ is used as the embedding for the missing modality and is subsequently fed into the pre-trained MMT. Additionally, the aforementioned process is applied to scenarios involving missing images to obtain the corresponding vision patch embedding $\hat{\mathcal{V}}_i$.

Dynamic Prompt-Tuning Given a pre-trained MMT f_θ with N consecutive Multi-head Self-Attention (MSA) layers, we denote the input representation of the b-th MSA layer as $\mathbf{h}^b \in \mathbb{R}^{L\times d}$, $b = 1, 2, \ldots, N$ with input length L and embedding dimension d. For full-modality data, we utilize the embedding layer of the pre-trained model $f_\theta(\cdot)$ to obtain the corresponding text embedding $\mathbf{E}^t$ and image embedding $\mathbf{E}^v$. In the case of missing-modality, we employ the generated word embedding $\hat{\mathcal{W}}$ and vision patch embedding $\hat{\mathcal{V}}$ to fill the corresponding missing modality. $\mathbf{h}^1$ is the concatenation of text embedding $\mathbf{E}^t$ and image embedding $\mathbf{E}^v$. The context-aware prompts $\mathbf{P}^t$, $\mathbf{P}^v$, and $\mathbf{P}^l$ are then attached to the embedding features along the sequence-length dimension to form the extended features $\mathbf{h}_p^b = [\mathbf{P}^t, \mathbf{P}^v, \mathbf{P}^l, \mathbf{h}^b]$. These extended features $\mathbf{h}_p^b$ are fed into the MMT starting from the b-th layer and continue to propagate through the remaining layers. The final output $\mathbf{h}_p^N$ represents comprehensive modal representation after the N-th layer. Rather than adding prompts at each MSA layer, which can result in considerable overhead, we selectively insert the prompts into the specific b-th layer.

Label-Augmented Prediction To further leverage the contextual information in label-level prompts, we design a label-augmented classifier by computing the similarity between the output representation of the MMT and the label matrix $\tilde{\mathbf{P}}^l$. Specifically, for the final prediction, we feed the output representation $\mathbf{h}_p^N$ into the pooler layer to obtain the representation $\boldsymbol{Z} \in \mathbb{R}^{d\times 1}$. Next, we calculate the probabilities

Missing Type	MM-IMDb						HateMemes			Food101		
	Text		Image		Both		Text	Image	Both	Text	Image	Both
Method	F1-M	F1-S	F1-M	F1-S	F1-M	F1-S	AUROC	AUROC	AUROC	ACC	ACC	ACC
SMIL	38.32	38.55	27.57	35.27	35.12	31.87	50.32	58.50	54.63	51.83	49.86	46.77
TFR-Net	37.70	38.82	38.14	39.45	37.24	38.11	51.18	55.57	52.12	65.91	67.58	63.41
AcMAE	47.47	46.73	43.82	42.20	44.05	43.75	55.74	59.66	57.25	69.28	73.75	71.15
IF-MMIN	39.63	38.10	31.95	26.89	31.98	29.33	57.62	53.44	55.19	66.76	64.36	68.53
ShaSpec	44.04	42.05	44.23	42.53	44.06	42.13	58.75	60.30	<u>60.96</u>	60.99	74.87	70.02
DrFuse	47.05	45.22	43.58	42.19	<u>48.83</u>	<u>47.15</u>	57.60	<u>60.66</u>	55.84	66.30	75.09	68.23
CorrKD	44.82	45.27	39.48	39.11	41.20	40.51	58.74	55.59	57.91	61.37	66.83	62.87
MAPs	46.12	45.47	<u>44.86</u>	<u>43.19</u>	45.48	44.30	58.62	60.16	58.89	67.02	75.62	72.52
MSPs	<u>49.16</u>	<u>48.81</u>	44.62	43.06	48.28	46.71	<u>59.60</u>	60.05	59.08	<u>71.74</u>	<u>79.09</u>	<u>74.46</u>
RAGPT	**55.16**	**55.00**	**46.44**	**45.12**	**50.89**	**50.22**	**64.10**	**62.57**	**63.47**	**75.53**	**81.98**	**76.94**
Improv. (%)	12.21↑	12.68↑	3.52↑	4.47↑	4.22↑	6.51↑	7.55↑	3.15↑	4.12↑	5.28↑	3.65↑	3.33↑
p-val.	$8.93e^{-6}$	$1.73e^{-5}$	$5.94e^{-5}$	$9.68e^{-6}$	$6.43e^{-6}$	$2.92e^{-5}$	$1.24e^{-6}$	$3.44e^{-5}$	$1.03e^{-5}$	$1.63e^{-6}$	$3.24e^{-6}$	$8.50e^{-5}$

Table 1: Performance comparison on three datasets with a 70% missing rate across various missing-modality scenarios. The best results are in **bold** font and the second <u>underlined</u>. Higher values of F1-M, F1-S, AUROC, and ACC indicate better performance.

Dataset	# Image	# Text	# Train	# Val	# Test
MM-IMDb	25,959	25,959	15,552	2,608	7,799
HateMemes	10,000	10,000	8,500	500	1,500
Food101	90,688	90,688	67,972	-	22,716

Table 2: Statistics of three multimodal downstream datasets.

$\hat{\mathbf{y}} \in \mathbb{R}^{C\times 1}$ for C classes: $\hat{\mathbf{y}} = \text{softmax}(\tilde{\mathbf{P}}^l * \boldsymbol{Z})$. During training, we freeze all parameters in the MMT and optimize the model using cross-entropy loss.

Experiments

Experimental Settings

A summary of the experimental settings is provided in this section, which refers to datasets, baselines, evaluation metrics, setting of missing pattern, and implementation details.

Datasets Following previous work (Lee et al. 2023; Jang, Wang, and Kim 2024), we evaluate our RAGPT on three downstream tasks. (1) MM-IMDb (Arevalo et al. 2017), primarily used for movie genre classification involving both image and text modalities. (2) Food101 (Wang et al. 2015), which focuses on image classification that incorporates both image and text. (3) HateMemes (Kiela et al. 2020), aimed to identify hate speech in memes using image and text modalities. Detailed statistics of datasets are presented in Table 2. The dataset splits are consistent with the original paper.

Baselines We compare our RAGPT with 9 competitive baselines, which are classified into three categories: (1) *Cross-modal generation methods*: SMIL (Ma et al. 2021), TFR-Net (Yuan et al. 2021), and AcMAE (Woo et al. 2023). (2) *Joint learning methods*: IF-MMIN (Zuo et al. 2023), ShaSpec (Wang et al. 2023), DrFuse (Yao et al. 2024), and CorrKD (Li et al. 2024). (3) *Prompt-based methods*: MAPs (Lee et al. 2023) and MSPs (Jang, Wang, and Kim 2024).

Evaluation Metrics Following prior works (Lee et al. 2023; Jang, Wang, and Kim 2024), we adopt appropriate dataset-specific metrics for evaluation: F1-Micro (F1-M) and F1-Sample (F1-S) for the MM-IMDb dataset, AUROC for the HateMemes dataset, and classification accuracy (ACC) for the Food101 dataset.

Setting of Missing Pattern We define the missing rate $\eta\%$ as the proportion of modality-incomplete data relative to the entire dataset. For each dataset, there are three possible cases of missing-modality: text missing, image missing, and both modalities missing. Text/image missing with a missing rate of $\eta\%$ indicates that there are $\eta\%$ instances consisting of texts/images and (1-η)% instances that contain both modalities. Missing both modalities with a missing rate of $\eta\%$ indicates that there are $\frac{\eta}{2}\%$ instances consisting solely of images, $\frac{\eta}{2}\%$ instances consisting solely of text, and (1-η)% instances that are complete, containing both modalities.

Implementation Details Following prior works (Lee et al. 2023; Jang, Wang, and Kim 2024), we utilize the pre-trained ViLT (Kim, Son, and Kim 2021) as our MMT backbone. The memory $\mathcal{B}$ for each dataset is constructed with the corresponding training set. The length l of context-aware prompts is set to 2, the number of retrieved instances K is chosen from $\{1, 3, 5, 7, 9\}$, and the prompt insertion layer b is set to 2. We utilize the AdamW optimizer (Loshchilov and Hutter 2017) with a learning rate of 1×10^{-3} and total 20 epochs for optimizing the parameters. All experiments are conducted with an NVIDIA RTX 3090 GPU.

Overall Performance

To verify the superiority of RAGPT, we compare it with 9 competitive baselines on three datasets under a missing rate of $\eta\% = 70\%$. From these results, we have the following observations:

First, our RAGPT consistently outperforms all strong baselines on three datasets under various modal conditions and metrics. Moreover, we retrain RAGPT and the best-performing baseline five times to calculate the p-value. Notably, RAGPT achieves improvements of 12.21% and 12.68% in the F1-M and F1-S metrics, respectively, on the MM-IMDb dataset with missing text. These results validate our design of exploiting expressive knowledge from

Module	Variant	MM-IMDb F1-M	MM-IMDb F1-S	HateMemes AUROC	Food101 ACC
RAGPT	**All**	**55.16**	**55.00**	**64.10**	**75.53**
Retriever	CM Retriever	52.37	51.70	61.87	74.24
	w/o Retriever	49.25	48.36	60.29	73.60
Generator	Padding	51.14	51.63	61.30	72.78
	w/o Filter	54.15	52.99	60.67	74.07
Prompter	Static Prompt	54.38	53.14	62.65	74.40
	w/o Label	53.41	53.45	62.01	74.34
	w/o Prompter	51.49	50.43	60.94	72.65

Table 3: Ablation study of RAGPT under 70% text missing.

retrieved instances to enhance both missing and complete modality data. Meanwhile, the missing modality generator and context-aware prompter distill expressive contextual information from retrieved instances to approximate missing content and generate dynamic prompts, respectively, thereby improving model robustness for incomplete modalities.

Second, *cross-modal generation* and *joint learning methods* demonstrate inferior performance, primarily due to the uncertainty introduced by random placeholders and the challenges of modality heterogeneity in reconstruction, which create significant performance bottlenecks. Moreover, *prompt-based methods* also exhibit limited effectiveness in missing-modality scenarios, as they rely on dummy imputations and static prompting strategies, further restricting their potential and resulting in performance stagnation.

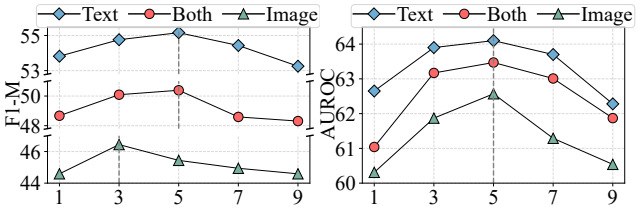
Ablation Study

We conduct various ablation experiments to evaluate the impact of each component within RAGPT under a 70% text missing case and summarize the results in Table 3.

Effect of Multi-Channel Retriever To analyze the impact of the retriever in RAGPT, we designed two variants: (1) **CM Retriever**: replacing the multi-channel retriever with cross-modal retriever, and (2) **w/o Retriever**: removing the retriever entirely. These results confirm the presence of the modal gap problem in cross-modal retrieval, which renders the retrieved instances irrelevant to the target images. Furthermore, this finding reinforces our design of the multi-channel retrieval that retrieves relevant instances by calculating within-modality similarities, thereby enhancing both missing and complete modality data.

Effect of Missing Modality Generator To evaluate the impact of the missing modality generator, we designed variant models: (1) **Padding**: using random values to fill in the missing modality, and (2) **w/o Filter**: removing the filter block entirely. We observe that dummy padding results in a decline in performance. This finding supports our assertion that dummy padding contributes to performance bottlenecks in prompt-based methods. Additionally, the removal of the filter layer leads to a significant performance drop, underscoring the importance of the filter layer in RAGPT for effectively mitigating noise.

Effect of Context-Aware Prompter To analyze the context-aware prompts, we design variants: (1) **Static Prompt**: replacing context-aware prompts with static prompts; (2) **w/o Label**: removing label enhancement; and (3) **w/o Prompter**: eliminating text-, vision-, label-prompts entirely. The three variants result in poorer performance, validating that static prompts offer limited relevant cues for addressing incomplete multimodal learning.



(a) Effect of K on MM-IMDb (b) Effect of K on HateMemes

Figure 3: Hyper-parameter analysis of K under three modality missing scenarios.

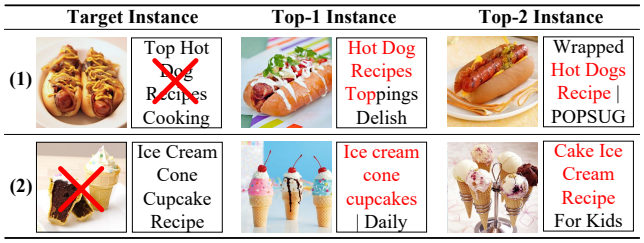


Figure 4: Examples of Top-2 retrieved instances for two modality-incomplete target instances. The first target instance is image-only while the second one is text-only. Red texts highlight similar content.

Hyper-Parameter Analysis

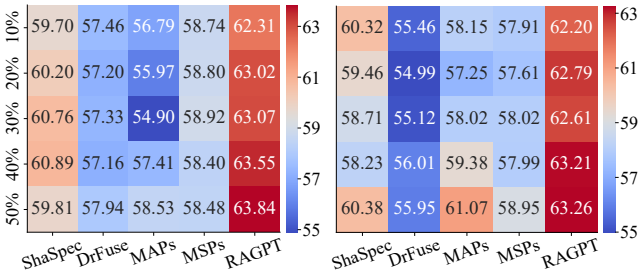
Fig. 3(a) and 3(b) present the sensitivity analysis of RAGPT's hyper-parameters K on the MM-IMDb and HateMemes datasets. The results demonstrate that the performance of RAGPT is improved by retrieving relevant instances. However, incorporating a larger number of instances may result in a decline in performance due to the introduction of noise (i.e., irrelevant instances). Consequently, we adopt $K = 3$ under the image missing case on the MM-IMDb dataset and $K = 5$ under other scenarios.

Retrieval Quality Presentation

To further analyze the efficacy of our proposed multi-channel retriever, we randomly select two instances with incomplete modalities from the Food101 dataset. Fig. 4 visualizes the Top-2 similar retrieved instances, demonstrating a strong semantic correlation between the retrieved and target instances in both image and text modalities. The high quality of retrieval relevance indicates our multi-channel retriever's ability to effectively identify relevant modal information.

Model Generalizability

To investigate the model's generalizability, we design two experiments with varying missing rates in the training set



(a) Text Missing (b) Both Missing

Figure 5: Generalization analysis on the HateMemes dataset across various missing rates in terms of AUROC.

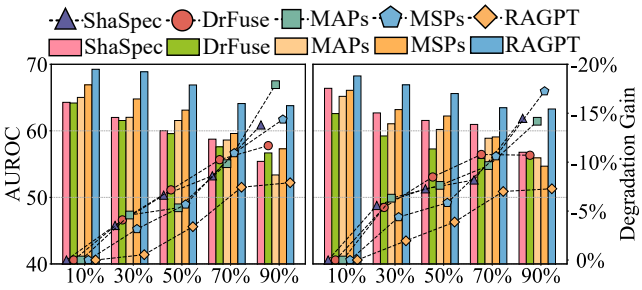


Figure 6: Robustness analysis on the HateMemes dataset across various missing rates in terms of AUROC.

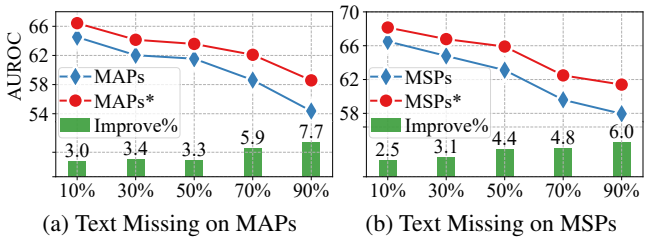
and evaluate their performance on a test set with a 90% missing rate. Compared with four strong baselines (ShaSpec, DrFuse, MAPs, and MSPs), Fig. 5(a) shows the results for the missing-text case, while Fig. 5(b) presents the results for scenarios of missing both modalities. We observe that our RAGPT outperforms all baselines across all missing rates, demonstrating superior performance for missing-modality. These results highlight RAGPT's generalizability, which can be attributed to the ability of exploring crucial cues from relevant contexts.

Robustness to Different Missing Rates

We conduct an experiment to analyze the model's robustness to varying missing rates. Fig. 6 illustrates the results comparing RAGPT with four strong baselines (ShaSpec, DrFuse, MAPs, and MSPs) on the HateMemes dataset. We observe that the performance of all baselines deteriorates markedly as the missing rate increases. In contrast, RAGPT demonstrates only a slight performance decrease as the missing rate increases. This result highlights the valuable components of RAGPT for effectively mitigating the impact of missing data. Specifically, RAGPT leverages expressive knowledge from retrieved instances to approximate missing modalities through the missing modality generator. Additionally, RAGPT generates context-aware prompts that enhance the performance of the pre-trained MMTs.

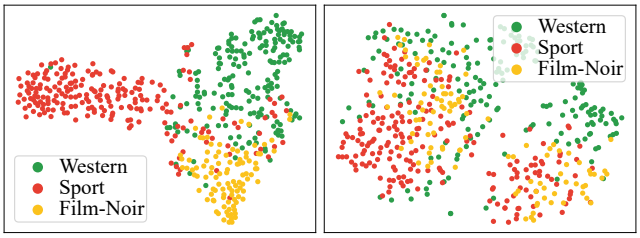
Model Scalability

To further validate the RAGPT's scalability, we integrate key modules (multi-channel retriever, missing modality generator, and context-aware prompter) into two prompt-based baselines (MAPs and MSPs). In Fig. 7, we observe a significantly slower rate of performance decline in the two baselines as the missing rate increased. This finding indicates that our modules significantly enhance the robustness of these baselines for incomplete modalities. It also validates the effectiveness of our design in extracting informative multimodal cues from relevant instances and prompting pre-trained MMTs.



(a) Text Missing on MAPs (b) Text Missing on MSPs

Figure 7: Effect of integrating key modules in RAGPT for baselines on the HateMemes dataset in terms of AUROC.



(a) RAGPT (b) MSPs

Figure 8: t-SNE visualization of RAGPT and MSPs on the MM-IMDb dataset under a 90% text missing rate.

Model Prediction Visualization

Fig. 8 illustrates the t-SNE (Van der Maaten and Hinton 2008) visualization of the embedding distributions for three genres (i.e., Sport, Film-Noir, and Western) in the MM-IMDb test set under a 90% text missing rate. We observe that while baseline MSPs learns distinguishable features, the learned features remain intertwined. In contrast, the representations of three genres learned by our RAGPT are more discriminative, exhibiting larger segregated areas among instances with different labels.

Conclusion

In this work, we proposed RAGPT, a novel retrieval-augmented dynamic prompt-tuning framework to address the missing-modality issue. This model-agnostic framework includes three key components: (1) the multi-channel retriever, (2) the missing modality generator, and (3) the context-aware prompter, to effectively inject valuable contextual knowledge into pre-trained MMT, thereby enhancing its robustness in the missing-modality scenario. Extensive experiments conducted on three real-world datasets demonstrate the superiority of RAGPT in tackling incomplete modality learning.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No.62176043, No.62072077, and No.U22A2097), and Kashgar Science and Technology Bureau (Grant No.KS2023025).

References

Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; and González, F. A. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Cheng, Z.; Ye, W.; Liu, L.; Tai, W.; and Zhou, F. 2023. Enhancing Information Diffusion Prediction with Self-Supervised Disentangled User and Cascade Representations. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 3808–3812.

Cheng, Z.; Zhang, J.; Xu, X.; Trajcevski, G.; Zhong, T.; and Zhou, F. 2024a. Retrieval-augmented hypergraph for multimodal social media popularity prediction. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 445–455.

Cheng, Z.; Zhou, F.; Xu, X.; Zhang, K.; Trajcevski, G.; Zhong, T.; and Philip, S. Y. 2024b. Information Cascade Popularity Prediction via Probabilistic Diffusion. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

Ghosh, A.; Acharya, A.; Jain, R.; Saha, S.; Chadha, A.; and Sinha, S. 2024. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 22031–22039.

Hodges, N. J.; Williams, A. M.; Hayes, S. J.; and Breslin, G. 2007. What is modelled during observational learning? *Journal of Sports Sciences*, 25(5): 531–545.

Jang, J.; Wang, Y.; and Kim, C. 2024. Towards Robust Multimodal Prompting with Missing Modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8070–8074.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19113–19122.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems (Neurips)*, 33: 2611–2624.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning (ICML)*, 5583–5594. PMLR.

Lee, H.-C.; Lin, C.-Y.; Hsu, P.-C.; and Hsu, W. H. 2019. Audio feature generation for missing modality problem in video action recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3956–3960.

Lee, Y.-L.; Tsai, Y.-H.; Chiu, W.-C.; and Lee, C.-Y. 2023. Multimodal prompting with missing modalities for visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14943–14952.

Li, M.; Yang, D.; Zhao, X.; Wang, S.; Wang, Y.; Yang, K.; Sun, M.; Kou, D.; Qian, Z.; and Zhang, L. 2024. Correlation-Decoupled Knowledge Distillation for Multimodal Sentiment Analysis with Incomplete Modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12458–12468.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 4582–4597.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. *arXiv e-prints*, arXiv–1711.

Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; and Peng, X. 2022. Are multimodal transformers robust to missing modality? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18177–18186.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 2302–2310.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763. PMLR.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Wang, H.; Chen, Y.; Ma, C.; Avery, J.; Hull, L.; and Carneiro, G. 2023. Multi-modal learning with missing modality via shared-specific feature modelling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15878–15887.

Wang, H.; Luo, S.; Hu, G.; and Zhang, J. 2024. Gradient-Guided Modality Decoupling for Missing-Modality Robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 15483–15491.

Wang, X.; Kumar, D.; Thome, N.; Cord, M.; and Precioso, F. 2015. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia & Expo Workshops (ICME)*, 1–6.

Woo, S.; Lee, S.; Park, Y.; Nugroho, M. A.; and Kim, C. 2023. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, 2776–2784.

Xu, X.; Zhou, F.; Zhang, K.; Liu, S.; and Trajcevski, G. 2021. Casflow: Exploring hierarchical structures and propagation uncertainty for cascade prediction. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 35(4): 3484–3499.

Yao, W.; Yin, K.; Cheung, W. K.; Liu, J.; and Qin, J. 2024. DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 16416–16424.

Yuan, Z.; Li, W.; Xu, H.; and Yu, W. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 4400–4407.

Zhao, J.; Li, R.; and Jin, Q. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2608–2618.

Zhong, T.; Lang, J.; Zhang, Y.; Cheng, Z.; Zhang, K.; and Zhou, F. 2024. Predicting Micro-video Popularity via Multimodal Retrieval Augmentation. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 9–16.

Zhou, F.; Xu, X.; Trajcevski, G.; and Zhang, K. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)*, 54(2): 1–36.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, K.; Yu, H.; Zhao, W. X.; and Wen, J.-R. 2022b. Filter-enhanced MLP is all you need for sequential recommendation. In *Proceedings of the ACM Web Conference (WWW)*, 2388–2399.

Zuo, H.; Liu, R.; Zhao, J.; Gao, G.; and Li, H. 2023. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.