Multimodal Prompting with Missing Modalities for Visual Recognition

Yi-Lun Lee[†] Yi-Hsuan Tsai[‡] Wei-Chen Chiu[†] Chen-Yu Lee[‡]
[†]National Yang Ming Chiao Tung University [‡]Google
{yllee10727, walon}@cs.nctu.edu.tw, {yhtsai, chenyulee}@google.com

Abstract

In this paper, we tackle two challenges in multimodal learning for visual recognition: 1) when missing-modality occurs either during training or testing in real-world situations; and 2) when the computation resources are not available to finetune on heavy transformer models. To this end, we propose to utilize prompt learning and mitigate the above two challenges together. Specifically, our modality-missing-aware prompts can be plugged into multimodal transformers to handle general missing-modality cases, while only requiring less than 1% learnable parameters compared to training the entire model. We further explore the effect of different prompt configurations and analyze the robustness to missing modality. Extensive experiments are conducted to show the effectiveness of our prompt learning framework that improves the performance under various missing-modality cases, while alleviating the requirement of heavy model re-training. Code is available.[1]

1. Introduction

Our observation perceived in daily life is typically mulitmodal, such as visual, linguistic, and acoustic signals, thus modeling and coordinating multimodal information is of great interest and has broad application potentials. Recently, multimodal transformers [13, 17, 22, 25, 35] emerge as the pre-trained backbone models in several multimodal downstream tasks, including genre classification [22], multimodal sentiment analysis [25, 35], and cross-modal retrieval [13,15,17,30], etc. Though providing promising performance and generalization ability on various tasks, there are still challenges for multimodal transformers being applied in practical scenarios: 1) how to efficiently adapt the multimodal transformers without using heavy computation resource to finetune the entire model? 2) how to ensure the robustness when there are missing modalities, e.g., incomplete training data or observations in testing?

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

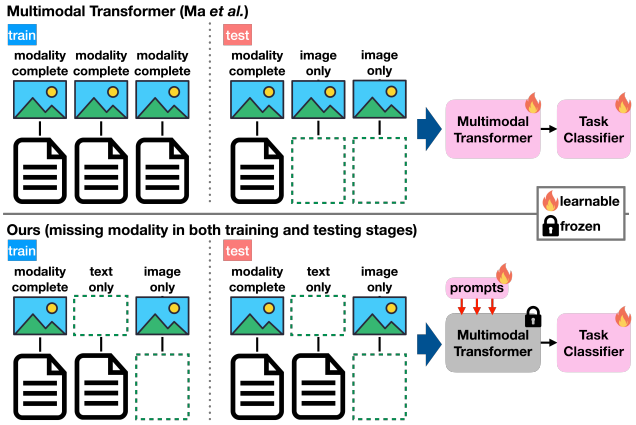


Figure 1. Illustration of missing-modality scenarios in training multimodal transformers. Prior work [22] investigates the robustness of multimodal transformers to modality-incomplete test data, with the requirement to finetune the entire model using modality-complete training data. In contrast, our work studies a more general scenario where various modality-missing cases would occur differently not only for each data sample but also learning phases (training, testing, or both), and we adopt prompt learning to adapt the pre-trained transformer for downstream tasks without requiring heavy computations on finetuning the entire model.

Most multimodal transformer-based methods have a common assumption on the data completeness, which may not hold in practice due to the privacy, device, or security constraints. Thus, the performance may degrade when the data is modality-incomplete (regardless of training or testing). On the other hand, transformers pretrained on large-scale datasets are frequently adopted as backbone and finetuned for addressing various downstream tasks, thanks to the strong generalizability of transformers. However, as the model size of transformers increases (e.g., up to billions of parameters [5,26,27]), finetuning becomes significantly expensive (e.g., up to millions of A100-GPU-hours [31]) and is even not feasible for practitioners due to the limited computation resources in most real-world applications. In addition, finetuning a transformer on relatively small-scale target datasets can result in restricted generalizability [9, 10] and stability [24], thus hindering it from being reused for

further learning with new data or in other tasks.

This motivates us to design a method that allows multimodal transformers to alleviate these two real-world challenges. One pioneer work [22] investigates the sensitivity of vision-language transformers against the presence of modal-incomplete test data (i.e., either texts or images are missing). However, they only consider the case of missing a specific modality for all the data samples, while in real-world scenarios the missing modality for each input data could not be known in advance. Moreover, [22] introduces additional task tokens to handle different missing-modal scenarios (e.g., text-only token when missing visual modality) and requires to optimize cross-modal features in the model. Hence finetuning the entire transformer becomes inevitable, leading to significant computation expense.

In this paper, we study multimodal transformers under a more general modality-incomplete scenario, where various missing-modality cases may occur in any data samples, e.g., there can be both text-only and image-only data during training or testing. In particular, we also focus on alleviating the requirement of finetuning the entire transformers. To this end, we propose a framework stemmed from prompt learning techniques for addressing the aforementioned challenges. Basically, prompt learning methods [2,5,8,16,18,32,42] emerge recently as efficient and effective solutions for adapting pre-trained transformers to the target domain via only training very few parameters (i.e., prompts), and achieve comparable performance with finetuning the whole heavy model. As motivated by [29] which shows that prompts are good indicators for different distributions of input, we propose to regard different situations of missing modalities as different types of input and adopt the learnable prompts to mitigate the performance drop caused by missing modality. As a result, the size of our learnable prompts can be less than 1% of the entire transformer, and thus the computation becomes more affordable compared to holistic finetuning. The key differences between our work and [22] are illustrated in Figure 1.

In order to further explore the prompt designs for multimodal transformers to tackle the general modality-incomplete scenario, we investigate two designs of integrating our missing-aware prompts[2] into pre-trained multimodal transformers: 1) input-level, and 2) attention-level prompt learning. We find that, the location of attaching prompts to transformers is crucial for the studied missing-modality cases in this paper, which also aligns the findings in [36], though under a different problem setting.

We conduct experiments to explore different prompt configurations and have observations of the impact on the length and location of prompts: 1) As the number of prompting layers increases, the model performs better intuitively but it is not the most important factor; 2) Attaching prompts to the layers near the data input achieves better performance; 3) The prompts' length has slight impact on model performance for attention-level prompts but may influence input-level prompts more on certain datasets. Moreover, we show extensive results to validate the effectiveness of adopting our prompting framework to alleviate the missing-modality issue under various cases, while reducing the learnable parameters to less than 1% compared to the entire model. Our main contributions are as follows:

- We introduce a general scenario for multimodal learning, where the missing modality may occur differently for each data sample, either in training or testing phase.
- We propose to use missing-aware prompts to tackle the missing modality situations, while only requiring less than 1% parameters to adapt pre-trained models, thus avoiding finetuning heavy transformers.
- We further study two designs of attaching prompts onto different locations of a pretrained transformer, input-level and attention-level prompting, where the input-level prompting is generally a better choice but the attention-level one can be less sensitive to certain dataset settings.

2. Related Work

Missing-Modality for Multimodal Learning. Multimodal learning methods leverage the complementary property from different modalities (e.g., images, texts, or audio) for learning to describe a common concept. Recently, multimodal transformers emerge as unified models that process inputs from different modalities and fuse them via token concatenation without modality-specific feature extractors. Such transformer-based models are widely applied in various mutlimodal tasks [4, 7, 13, 17]. However, most multimodal learning methods require the completeness of modality, which may not be the real-world case. Once one of the modalities is missing, the multimodal fusion becomes unreachable and the model may predict inaccurately [22].

To this end, recent works [22,23,39,40] explore to build multimodal models which are robust to data with missing modalities. SMIL [23] is proposed to estimate the latent features of the modality-incomplete data via Bayesian Meta-Learning. Zeng *et al*. [39] propose a tag encoding module to assist the transformer's encoder learning with different missing modalities. MMIN [40] predicts the representation of any missing modality given other available modalities via learning a joint multimodal representations. Ma *et al*. [22] investigate the robustness of multimodal transformers to missing modalities and improve it via multitask optimization to achieve better fusion strategies respectively for different missing-modality cases. In contrast, this paper conducts a more thorough study on multimodal transformer's robustness where various modality-missing would

[2]In this paper, we use "missing-aware prompts" and "modality-missing-aware prompts" interchangeably.

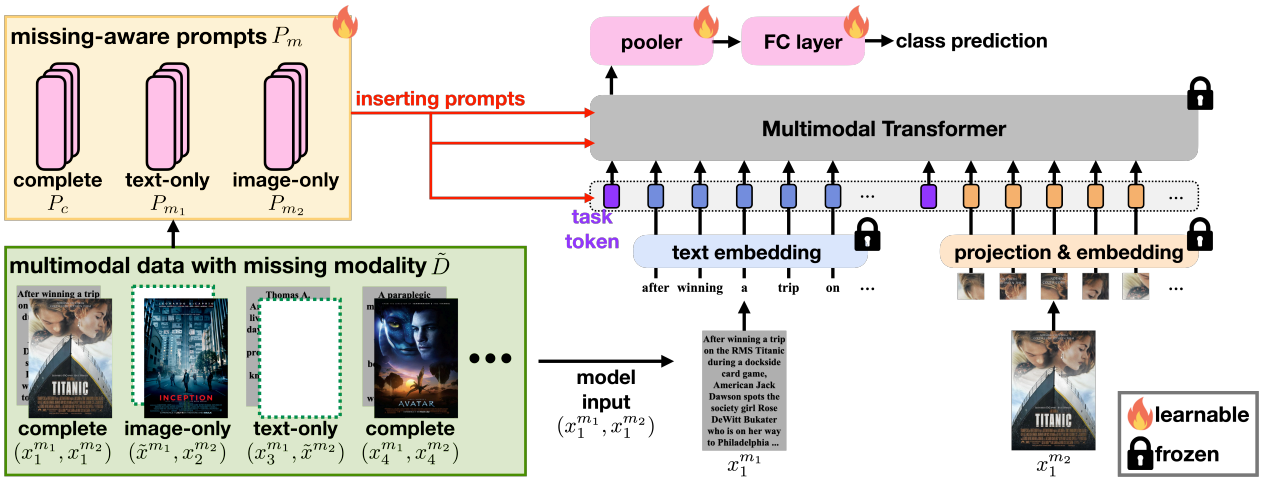


Figure 2. The overview of our proposed prompt-based multimodal framework. We first select the missing-aware prompts P_m according to the missing case (e.g., complete, text-only, image-only in vision-language tasks) of the multimodal inputs $(x_i^{m_1}, x_i^{m_2})$, in which the dummy inputs $\{\tilde{x}^{m_1}, \tilde{x}^{m_2}\}$ respectively for text and image are adopted for the corresponding missing modality. Then we attach missing-aware prompts into multiple MSA layers via different prompting approaches (see Figure 3 and Section 3.3). We select the text-related task token of the multimodal transformer as our final output features, and feed them to the pooler layer and fully-connected (FC) layers for class predictions. Note that only the pink-shaded blocks require to be trained while the others are frozen.

occur for any data samples and anywhere in learning phases, particularly focusing on reducing the computation of model finetuning.

Prompt Learning. Prompt learning emerges as an effective transfer learning technique in nature language processing, which adopts a "prompt" to modify the input text for instructing the pre-trained model for downstream tasks. For instance, with manually-chosen prompts, the pre-trained language models [5] show a strong generalizability to downstream tasks in few-shot or zero-shot manner. Instead of relying on human involvement to discover the proper prompts for adapting transformers into new tasks, prompt-tuning [16] and prefix-tuning [18] are proposed to automate the prompt learning in continuous space.

Recently, prompts are also introduced into computer vision tasks [2, 8, 36, 37] and multimodal learning tasks [11, 19, 32, 38, 41, 42]. The visual prompts [2, 8] are applied to the vision transformers and adapt the large-scale pre-trained model to downstream vision tasks via tuning very few learnable parameters. L2P [37] and DualPrompt [36] further adopt the prompts to learn different task information conditionally in continual learning. CoOp [42] models the context in prompts with learnable vectors in continuous space while the parameters of the entire CLIP-like pre-trained model are kept fixed. Frozen [22] encodes the image as a sequence of continuous embeddings to serve as a prefix prompt to instruct the pre-trained frozen language models in generating the appropriate caption by a multimodal few-shot learning manner. MaPLe [11] further applies prompts in both vision and language encoders to improve the alignment between vision and language representations. Besides, PromptFuse [19] utilized prompts to learn alignment among different modalities for parameter-efficient adaptation on downstream tasks. These works explore the great adaptation ability of prompt learning to different tasks with different input domains. This motivates us to integrate the prompt learning technique into multimodal learning under a general missing-modality scenario via regarding different missing cases as different learning tasks.

3. Proposed Method

3.1. Overall Framework

In this paper, we focus on multimodal learning with missing modalities in general situations. We assume that there are several missing-modality cases, e.g., missing one modality or missing more modalities, to represent the more realistic scenario of multimodal learning in the real world. Note that, the missing case during training can be also different from the one in testing. In addition, as the pretrained transformers become larger and untrainable with limited computation resources, it is crucial to develop the method without the need of finetuning the entire pretrained model.

Problem Definition. To be the simplest but without loss of generality, we consider a multimodal dataset consisting of $M = 2$ modalities m_1 and m_2 (e.g., image and text). Given a multimodal dataset $D = \{D^c, D^{m_1}, D^{m_2}\}$, we denote $D^c = \{x_i^{m_1}, x_i^{m_2}, y_i\}$ as the modality-complete subset, while $D^{m_1} = \{x_j^{m_1}, y_j\}$ and $D^{m_2} = \{x_k^{m_2}, y_k\}$ are denoted respectively as the modality-incomplete subsets (e.g., text-only and image-only) where one modality is missing. As shown in Figure 2, the training data may contain data

samples with different missing cases including complete data D^c, text-only data D^{m_1}, and image-only data D^{m_2}.

To preserve the format of multimodal inputs, we simply assign dummy inputs $\tilde{x}^{m_1}$, $\tilde{x}^{m_2}$ (e.g., empty string/pixel for texts/images) to the missing-modality data and obtain $\tilde{D}^{m_1} = \{x_j^{m_1}, \tilde{x}_j^{m_2}, y_j\}, \tilde{D}^{m_2} = \{\tilde{x}^{m_1}, x_k^{m_2}, y_k\}$. Therefore, the multimodal data with missing modality can be reformed as $\tilde{D} = \{D^c, \tilde{D}^{m_1}, \tilde{D}^{m_2}\}$.

For simplicity, we follow [22] to adopt the multimodal transformer ViLT [13] as our backbone model, which is pretrained on large-scale vision and language datasets. Note that the backbone model is untrainable in our scenario due to the limitation of computation resources. In order to tackle the missing modality, we propose **missing-aware prompts** to instruct the pretrained model's prediction with different input cases. These prompts are assigned according to the missing case of input data and attached to multiple blocks of the multimodal transformer. With the assumption of untrainable pretrained models, the only trainable parameters are the missing-aware prompts, pooler layer, and fully-connected layers for learning the multimodal classifier.

3.2. Prompt Learning for Missing Modalities

Prompt-based learning first emerges as the efficient method in natural language processing (NLP) for transfer learning without finetuning the whole pretrained model. In general, prompts are prepended to the input for instructing the model prediction. With a similar motivation, we propose missing-aware prompts to instruct the pretrained transformer, conditioned on different input cases of missing modality. To this end, we design the corresponding missing-aware prompts for each missing-modality case. As shown in Figure 2, we first assign $M^2 - 1$ prompts for M modality tasks (e.g., 3 missing-aware prompts for the vision-language task), and prepend them to the input according to the type of missing modality.

Given a pretrained multimodal transformer f_θ with N consecutive MSA layers, we denote the input embedding features of the i-th MSA layer as $h^i \in \mathbb{R}^{L \times d}, i = 1, 2, ..., N$ with input length L and embedding dimension d. Note that h^1 is the output of modality-specific embedding functions that pre-process the inputs to token sequences (i.e., BERT tokenizer for the text modality and visual embedding [6] layers for the image modality). Then the missing-aware prompts $p_m^i \in \mathbb{R}^{L_p \times d}$ are attached to the i-th layer, where L_p is the prompt length, d is the embedding dimension, and $m \in \{c, m_1, m_2\}$ represents different missing-modality cases. Finally, the missing-aware prompts are attached to the embedding features along with the input-length dimension to form extended features h_p^i:

$$h_p^i = f_{prompt}(p_m^i, h^i), \tag{1}$$

where f_{prompt} defines the approach to attach prompts to the embedding features, and will be detailed in the next section.

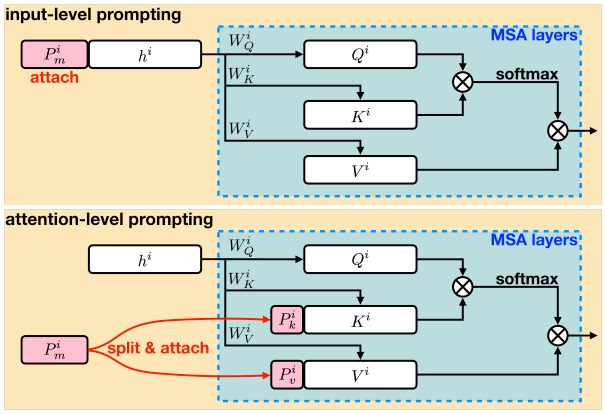


Figure 3. The illustration of two prompting approaches. The top block shows the input-level prompting method, which attaches the missing-aware prompts p_m^i to the input of the i-th MSA layer. The bottom block shows the attention-level prompting method, which first splits the missing-aware prompts p_m^i into two sub-prompts p_k^i, p_v^i with the same length, and attaches them respectively to the key K^i and value V^i in the i-th MSA layer (see Section 3.3).

Overall Objective. For model training, we freeze all the parameters f_θ of the multimodal transformer except for the task-specific layers f_{θ_t} (i.e., pooler layer and fully-connected layer), in order to output corresponding predictions based on each visual perception task. Moreover, we denote θ_p as the parameters of missing-aware prompts. The overall objective with trainable parameters is defined as:

$$L = L_{task}(x_i^{m_1}, x_i^{m_2}; \theta_t, \theta_p), \tag{2}$$

where $(x_i^{m_1}, x_i^{m_2}) \in \tilde{D}$ is the multimodal input pair with missing-modality cases, and L_{task} represents the task-specific multimodal objective, e.g., binary cross-entropy loss for movie genre classification.

3.3. Prompt Design

In this section, we focus on the design of the f_{prompt} function that attaches prompts to each input layer as in (1). In general, most prompt-based methods typically add the prompts to the input sequence and instruct the model for downstream tasks. However, [36] shows that the configuration of prompts and the position where prompts are added is crucial to prompt-based learning. In our situation, since the input modality may be missing, studying the proper configuration to attach prompts is of great importance. In Figure 3, we introduce two configurations of prompts: *input-level prompting* and *attention-level prompting*.

Input-level Prompting. A common approach to attach the prompts is to prepend prompts into input sequences for each layer, as shown in the top of Figure 3. The prompt function can be written as:

$$f_{prompt}^{input}(p_m^i, h^i) = [p_m^i; h^i], \tag{3}$$

where $[\cdots;\cdots]$ represents the concatenation operation. Assume there are N_p layers attaching the prompt parameters, the length of input/output sequence for each MSA layer would become larger as it goes deeper. For example, the length of sequence in the output of the last MSA layer with prompts would become $(N_pL_p + L)$. In this way, the prompts for the current layer can interact with the prompt tokens inheriting from previous layers, and thus learn more effective instructions for the model prediction. However, we find that such increasing length of input sequence makes the input-level prompt learning sensitive to the dataset with different multimodal token lengths, which may be less favorable in certain multimodal downstream tasks. We discuss the details in Section 4.2.

Attention-level Prompting. Another prompting approach is to modify the inputs of the MSA layers [33] with prompts. In the bottom of Figure 3, we split the prompt into two sub-prompts p_k^i, p_v^i with the same sequence length $\frac{L_p}{2}$ and prepend them to the key and value vectors respectively. We denote the query, key and value for the MSA layer as:

$$Q^i = h^i W_Q^i; K^i = h^i W_K^i; V^i = h^i W_V^i, \tag{4}$$

where $W_Q^i, W_K^i, W_V^i \in \mathbb{R}^{d\times d}$ is the projection weights for MSA layers. Then, we can define the prompt function for attention-level prompts as:

$$\begin{aligned} &f_{prompt}^{attn}(p_m^i, h^i) = \textsc{Attention}^i(p_m^i, h^i), \\ &\textsc{Attention}^i = softmax(\frac{Q^i[p_k^i, K^i]^T}{\sqrt{d}})[p_v^i; V^i]. \end{aligned} \tag{5}$$

The attention-level prompting provides another way to instruct the pretrained model from the perspective of the attention mechanism in transformers. As the prompts do not prepend to the query vector, the output sequence length remains the same as the input sequence.

Multi-layer prompting and locations where to attach prompts. Intuitively, different layers of the multimodal transformer have different context of feature embedding [28], and the effect of prompts for each layer may vary. With the self-attention mechanism, the input tokens from different modalities are fused closely along with the transformer layers. That is, the features of early layers may have more characteristics from different modalities than those of deeper layers that are well-fused to multimodal tokens with respect to the task objective. This motivates us to explore the most proper locations to attach missing-aware prompts.

Here, we introduce the multi-layer extension of prompts which can be denoted as $P_m = \{p_m^i\}_{i=start}^{end} \in \mathbb{R}^{N_p\times L_p\times d}$, where p_m^i is the prmopt attaching to the input sequence (input-level) or MSA layer (attention-level) of the i-th layer in transformers, and $N_p = (end - start + 1)$ is the total number of layers with prompts. Note that we simply assume that the chosen indices of MSA layers are continuous. Instead of attaching prompts to either whole layers or only the first layer, we empirically find that early half of layers is the best location starting from the first layer $(start = 0, end = \frac{N}{2} - 1)$ with $N_p = \frac{N}{2}$. More results and discussions are in Section 4.3.

4. Experimental Results

Datasets. We follow the work [22] to evaluate our methods on three multimodal downstream tasks:

• *MM-IMDb* [1] is a movie genre classification dataset with image and text modalities. As a movie might have several genres, the task is hence a multi-label classification that predicts the genre via using image, text, or both modalities.

• *UPMC Food-101* [34] is the classification dataset including image and text, in which it consists of noisy image-text paired data collected from Google Image Search and has identical categories to the largest publicly available ETHZ Food-101 dataset [3].

• *Hateful Memes* [12] is a more challenging multimodal dataset that aims to identify the hate speech in memes via using image and text modalities. To prevent the model from relying on a single modality, it is constructed to make unimodal models more likely to fail via adding challenging samples ("benign confounders") while the multimodal models are more likely to work better.

Metrics. As these datasets focus on different classification tasks, we use corresponding proper metrics for each dataset. For *MM-IMDb* [1], F1-Macro is adopted to measure the multi-label classification performance; For *UPMC Food-101* [34], the metric is the classification accuracy; For *Hateful Memes* [12], we use Area Underthe Receiver Operating Characteristic Curve (AUROC).

4.1. Implementation Details

Input. For the text modality, we use bert-base-uncased tokenizer to tokenize the text input. If the text is missing, we use an empty string as dummy input (i.e., $\tilde{x}^{m_1}$). The maximum length of text inputs varies according to the datasets: 1024 for MM-IMDb, 512 for UPMC Food-101, and 128 for Hateful Memes. For the image modality, we follow [13] to resize the shorter side of input images to 384 and limit the longer side to be under 640 while keeping the aspect ratio. Similar to [6], we decompose images into patches of size 32×32. If the image is missing, we create an image with all pixel values equal to one as dummy input (i.e., $\tilde{x}^{m_2}$).

Multimodal Backbone. We adopt the pre-trained multimodal transformer ViLT [13] as our backbone since it is commonly used in various transformer-based methods for learning multimodal tasks. ViLT stems from Vision Transformers [6] and advances to process multimodal inputs with the tokenized texts and patched images. Without using modality-specific feature extractors, ViLT is pre-

Datasets	Missing rate η	Training Image	Training Text	Testing Image	Testing Text	Baseline [13]	Attention-level prompts (Ours)	Input-level prompts (Ours)
MM-IMDb [1] (F1-Macro)	70%	100%	30%	100%	30%	35.13	38.16	**39.22**
		30%	100%	30%	100%	37.73	44.74	**46.30**
		65%	65%	65%	65%	36.26	41.56	**42.66**
Food101 [34] (Accuracy)	70%	100%	30%	100%	30%	66.29	72.57	**74.53**
		30%	100%	30%	100%	76.66	86.05	**86.18**
		65%	65%	65%	65%	69.25	78.09	**79.08**
Hateful Memes [12] (AUROC)	70%	100%	30%	100%	30%	60.78	**62.17**	59.11
		30%	100%	30%	100%	61.64	62.34	**63.06**
		65%	65%	65%	65%	62.48	64.55	**66.07**

Table 1. Quantitative results on the MM-IMDB [1], UPMC Food-101 [34], and Hateful Memes [12] datasets with missing rate $\eta\% = 70\%$ under various modality-missing scenarios. **Bold** number indicates the best performance.

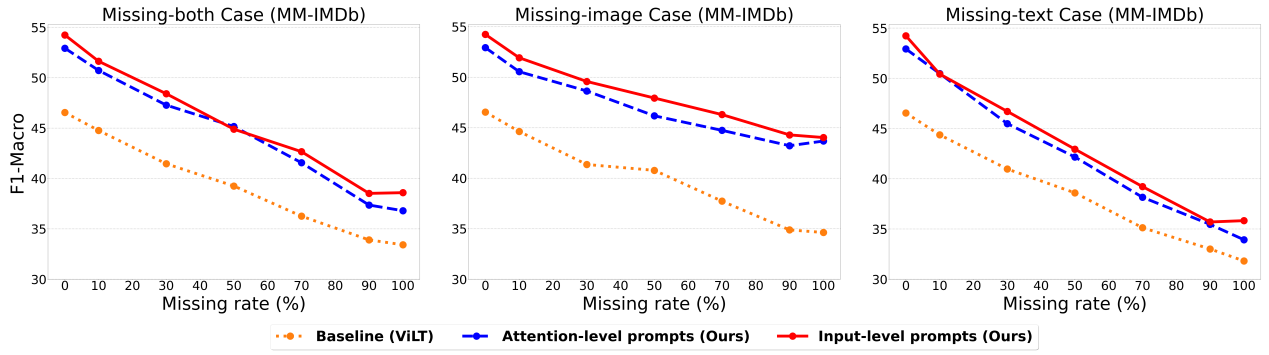


Figure 4. Quantitative results on the MM-IMDb dataset with different missing rates under different missing-modality scenarios. Each data point on the figure represents that training and testing are with the same $\eta\%$ missing rate.

trained on several large vision-language datasets (e.g., MS-COCO [20] and Visual Genome [14]) via objectives such as Image Text Matching and Masked Language Modeling.

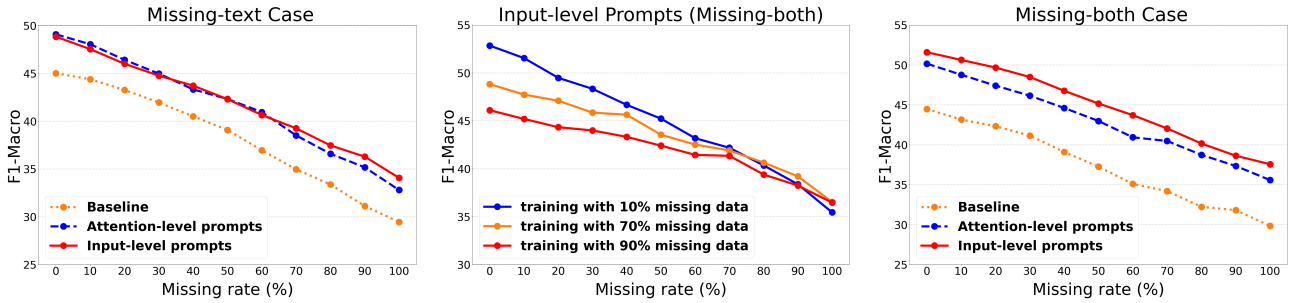
Model Training Details. We freeze all the parameters of the ViLT backbone to avoid heavy finetuning and only train the learnable prompts and parameters corresponding to downstream tasks (i.e., the pooler and task-specific classifier). The length L_p of learnable prompts is set to 16 by default, and the indices of MSA layers to attach prompts start from 0 and end at 5 (i.e., maximum 6 MSA layers are prompted). We use the AdamW optimizer [21] in all experiments. The base learning rate is 1×10^{-2} and weight decay is 2×10^{-2}. The learning rate is warmed up for 10% of the total training steps and then is decayed linearly to zero.

Setting of Missing Modality. We focus on the more general scenario where the missing modality can occur in both training and testing phases, where each modality for each data sample has chances to be lost. We define the missing rate $\eta\%$ as the proportion of modality-incomplete data to the entire dataset. In the vision and language tasks, there are three possible cases of missing modality: missing-text, missing-image, and missing-both. For the first two cases, missing-text (missing-image) with missing rate $\eta\%$ indicates that there are $\eta\%$ image-only (text-only) data and $(1-\eta)\%$ complete data. For the missing-both case, there are $\frac{\eta}{2}\%$ text-only data, $\frac{\eta}{2}\%$ image-only data, and $(1-\eta)\%$ complete data. Such partition can be extended to the tasks of M modality: having $(\frac{\eta}{M^2-2})\%$ modality-incomplete data for each missing case and $(1-\eta)\%$ complete data. In our experiments, the missing rate $\eta\%$ is set to 70% by default.

4.2. Main Results

We focus on studying the robustness of multimodal transformers against general incompleteness in multimodal data, without the requirement to finetune the entire heavy model. The main baseline to compare with is the model that only trains the pooler and the task-specific classifier, where the improved performance brought by our proposed method with respect to such baseline directly reflects the benefit from our missing-aware prompts.

In Table 1, we show quantitative results on three classification tasks with 1) various missing-modality cases, and 2) two prompt learning designs. First, we find that our attention-level prompting consistently improves the baseline by a large margin in all the scenarios, showing that our missing-aware prompts, without entire model finetuning, are able to tackle general missing-modality cases and provide a good instruction for model prediction. Next, we show that input-level prompting further improves the performance in most settings, except for one case on the Hateful Memes dataset, in which this sensitivity analysis will be discussed in the following section.



(a) Train: Missing-both 70%; Test: Missing-text (b) Train: Missing-both; Test: Missing-both (c) Train: Modality-complete; Test: Missing-both

Figure 5. Ablation study on robustness to the testing missing rate in different scenarios on MM-IMDb. (a) All models are trained on missing-both case with 70% missing rate, and evaluated on missing-text case with different missing rates. (b) Input-level prompts are trained on missing-both cases with 10%, 70%, and 90% missing rate, which represents more modality-complete data, balanced data, and less modality-complete data, respectively. Evaluation is on missing-both case with different missing rates. (c) All models are trained with modality-complete data, where each data pair can be randomly assigned with different missing modality at different training epochs (i.e., text-only, image-only, and modality-complete) to account for the possible missing modalities in the testing time. Evaluation is on missing-both case with different missing rates.

In Figure 4, we present more results from a wider range of missing rate on MM-IMDb. Similar trends can be concluded, where the input-level prompting shows favorable performance compared to the other two methods. More qualitative results, in which shows similar trends, are provided in the supplementary material due to limited space.

Performance Sensitivity. As introduced in Section 3.3, different configurations of prompts have distinct behaviors to learn the instruction for pre-trained models. Input-level prompting can learn more effective instructions for each prompting layer, with the information of their inherited prompt tokens from previous layers. However, the performance may become sensitive, depending on the prompt length and the input length of different datasets. For instance, the Hateful Memes dataset has shorter text sequences (i.e., 128) compared to other datasets, in which prompts may become more influential to the final performance if the prompt length is too long. This may cause ambiguity for the model to learn task-specific features, e.g., input-level prompting performs slightly worse on Hateful Memes, especially for the text-missing case in Table 1.

In contrast, attention-level prompts are attached to the key and value of each prompted MSA layer, and hence each prompt is only responsible for the instruction of its corresponding layer. While input-level prompting reaches the best performance in most cases, attention-level prompting can be more stable and less sensitive to different datasets. This observation indicates the trade-off in terms of model performance and stability for different prompt designs.

Efficiency on Parameters. In comparison with finetuning the entire model which needs to update 113M parameters of the ViLT, resulting in 46.45 F1-Macro score on MM-IMDb for the missing-both case, our proposed method keeps the pre-trained ViLT frozen and only needs to train the missing-aware prompts of 221K parameters (i.e., merely 0.2% compared to the entire model), while still achieving competitive performance of 42.66 in F1-Macro (see Table 1). Note that, the number of parameters for the task classifier is not taken into consideration as it is necessary for learning multimodal tasks. Particularly, once scaling up to the huge models with billions of parameters, our proposed missing-aware prompting would be more preferable and applicable for multimodal downstream tasks with missing modalities, with a better balance between computational cost and performance.

4.3. Ablation Study

Robustness to different missing rates. We conduct further experiments to analyze the robustness of our proposed method against different missing-modality rates between training and testing phases. We first evaluate the models that are trained on the missing-both case with missing rate 70%, and test with different missing-text situations. As shown in Figure 5(a), we find that both attention-level and input-level promptings are robust to different missing rates in testing time, comparing with the baseline. Moreover, we observe the trend that attention-level prompting performs better than the input-level one when testing with a missing rate lower than 30%, but gradually become less effective when the missing rate goes higher than 30%. Such a trend indicates that attention-level prompting learns better on modality-complete data, while input-level prompting is more robust to modality-incomplete data.

Moreover, in Figure 5(b), we examine three input-level prompting models trained with missing-both rates 10%, 70%, and 90% to represent the degree of missing situations during training: more modality-complete data (90% complete), balanced data (30% complete, 35% only-text, and 35% only-image), and less modality-complete data (10% complete) respectively. We find that when training with more modality-complete (i.e., low missing rate) data, the

performance is much higher when also testing with low missing rate. Interestingly, when testing with a high missing rate, the model trained with 90% incomplete data performs competitively even with model trained with much more complete data. This shows the robustness of our prompt design to handle different missing-modality cases.

Results with complete training data. In this paper, we assume that it is not guaranteed to collect entire modality-complete data due to privacy and budget limits in the real-world scenario. However, there still exists publicly available modality-complete datasets which are well annotated for training. Therefore, we conduct experiments with a new baseline and our methods by training with modality-complete data. To account for the missing-modality case during testing, we randomly select data with different modality-missing cases for each data pair (i.e., text-only, image-only, or complete). Note that, different from other experimental settings introduced in the paper, here one data pair can be in various missing-modality cases at different training epochs. The results are shown in Figure 5(c). Similarly, we find that our method consistently improves the baseline under different missing rates.

The effect of selected layers $(start, end)$**.** We conduct experiments to analyze the effect of locations to attach prompting layers in Figure 6. We observe that the performance increases intuitively as the number of layers increases, while a more critical factor is which layer to start attaching prompts, i.e., the earlier layer the better. This indicates that early layers with prompts influence model predictions more. One reason can be: multimodal inputs are fused from the beginning of the transformer, in which the degree of fusion increases when the layer is deeper (i.e., the characteristics of each modality remain more distinct in earlier layers). Therefore, it is more effective for early layers to be guided by the instruction of missing-aware prompts, before each modality loses its distinct characteristics.

The effect of prompt length L_p**.** We study the influence of prompt length in Figure 7. Intuitively, the performance is improved as the prompt length becomes longer. However, both prompting methods reach the best performance with the prompt length equal to 16, showing that the length should be balanced. In addition, to validate the efficiency of our method, we calculate the proportion of prompt parameters to parameters of the entire pre-trained model (numbers above the red data points in Figure 7. Our prompt-based method only requires an additional 0.2% parameters but improves baselines by a large margin. Even with fewer parameters (i.e., reducing the prompt length to 1), the performance is still competitive. We also conduct a new baseline with additional parameters in the task classifier of the same proportion, i.e., 0.2% of the entire model parameters, but it does not show clear improvement. It demonstrates the efficacy of the proposed method in learning multimodal tasks to handle missing-modality, only with very few trainable parameters.

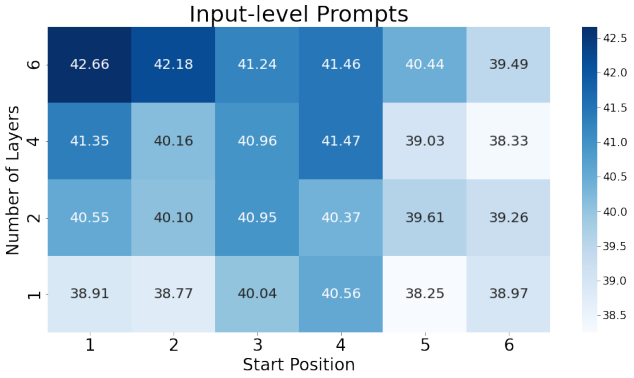


Figure 6. Ablation study on the location of prompting layers for input-level prompts.

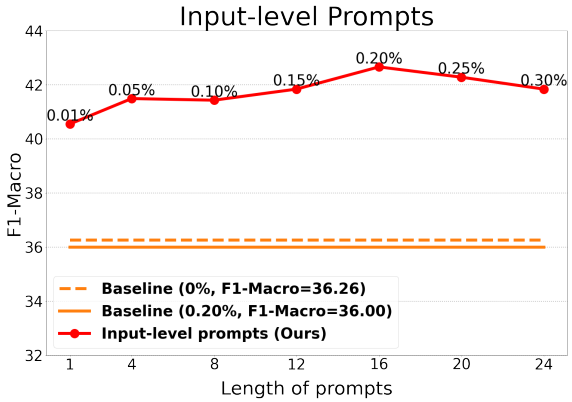


Figure 7. Ablation study on different length L_P of prompts for input-level prompts. The numbers above the red points are the proportion of parameters in prompts, compared to the entire model. We further conduct the new baseline with additional parameters with the same proportion (e.g., 0.2%) of the prompt size, denoted as the orange solid line.

5. Conclusions

We tackle two major challenges in multimodal learning: 1) a general scenario for missing modality that occurs either during training or testing, and 2) heavy computational requirement for training transformers. As a simple yet effective approach, we propose a missing-aware prompting method which is easy to plug in the transformer-like multimodal model to alleviate the performance drop caused by missing modality, while also not requiring heavy model finetuning. We further explore the configuration of prompts and show the robustness to the missing modalities during various scenarios. Extensive experiments and ablation studies demonstrate the effectiveness of our approach.

Acknowledgement. This work is supported by NSTC (National Science and Technology Council, Taiwan) 111-2628-EA49-018-MY4 and 111-2636-E-A49-003.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *International Conference on Learning Representations (ICLR) Workshops*, 2017. 5, 6

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *ArXiv:2203.17274*, 2022. 2, 3

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014. 5

[4] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 4, 5

[7] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[8] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[9] Haotian Ju, Dongyue Li, and Hongyang R Zhang. Robust fine-tuning of deep neural networks with hessian-based generalization guarantees. *ArXiv:2206.02659*, 2022. 1

[10] Mayank Kejriwal and Ke Shen. Do fine-tuned commonsense language models really generalize? *ArXiv:2011.09159*, 2020. 1

[11] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *ArXiv:2210.03117*, 2022. 3

[12] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5, 6

[13] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1, 2, 4, 5, 6

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 2017. 6

[15] Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. Formnet: Structural encoding beyond sequential modeling in form document information extraction. In *ACL*, 2022. 1

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *ArXiv:2104.08691*, 2021. 2, 3

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

[18] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *ArXiv:2101.00190*, 2021. 2, 3

[19] Sheng Liang, Mengjie Zhao, and Hinrich Schütze. Modular and parameter-efficient multimodal fusion with prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. 3

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014. 6

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2018. 6

[22] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 5

[23] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[24] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[25] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 1

[26] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv:2112.11446*, 2021. 1

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 2020. 1

[28] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

[29] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Prefix conditioning unifies language and label supervision. *ArXiv:2206.01125*, 2022. 2

[30] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, 2023. 1

[31] Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Bideman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. What language model to train if you have one million gpu hours? *ArXiv:2210.15424*, 2022. 1

[32] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

[34] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia & Expo (ICME) Workshops*, 2015. 5, 6

[35] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *ACM Web Conference (WWW)*, 2020. 1

[36] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 4

[37] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[38] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *ACM Conference on Multimedia (MM)*, 2022. 3

[39] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. *ArXiv:2204.13707*, 2022. 2

[40] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021. 2

[41] Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 3

[42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 2, 3