南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

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

CVPR 2023

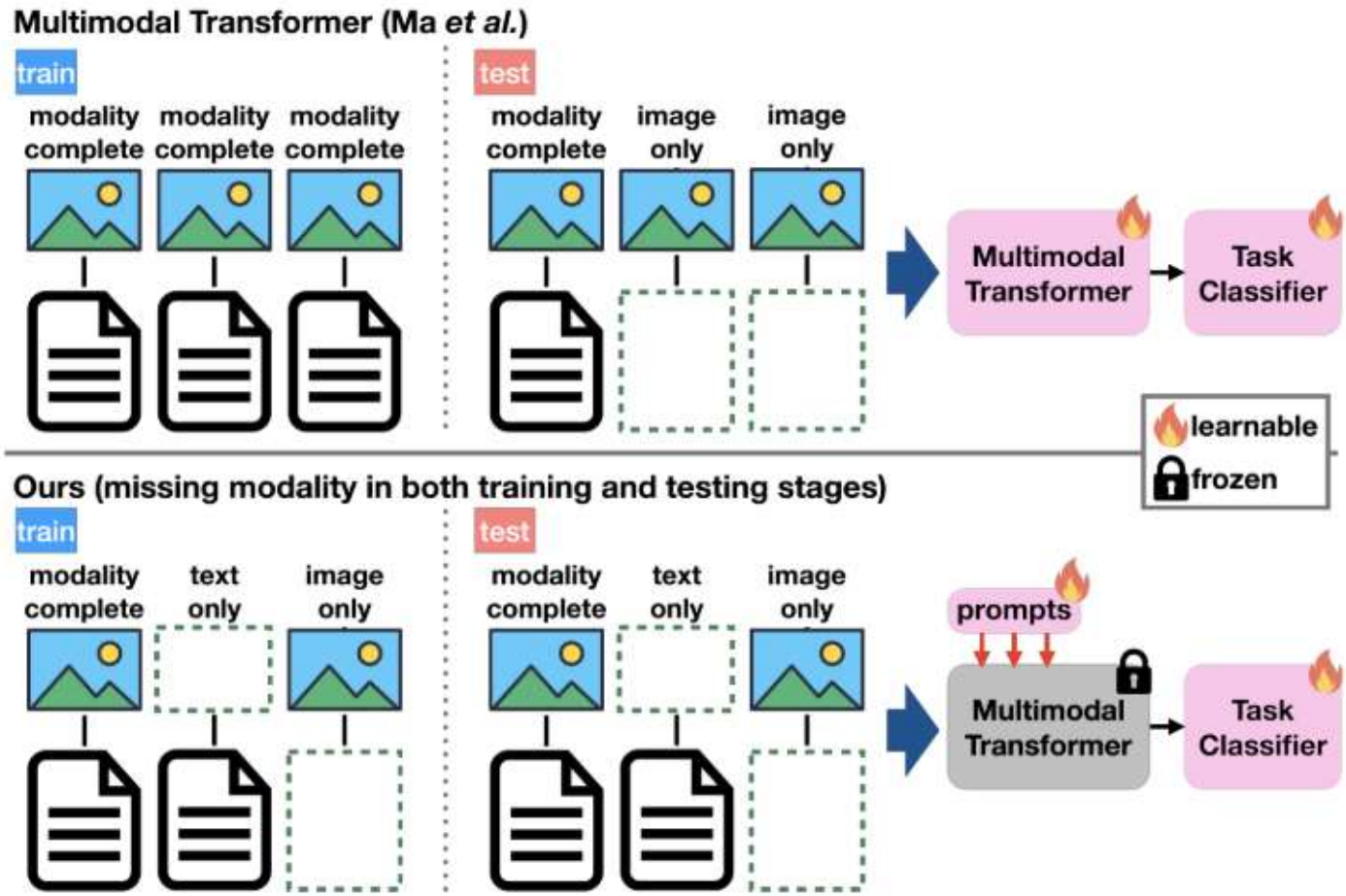
2024.7.15 朱新佳

Motivation

- 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.

- 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, finetuning becomes significantly **expensive** and is even not feasible for practitioners due to the limited computation resources in most realworld applications.

Introduction



missing-modality scenarios

- data sample
- learning phases

  (training, testing, or both)

training cost

prompt learning method

Problem Definition

- To be the simplest but without loss of generality, the author considers a multimodal dataset consisting of $M = 2$ modalities m_1 and m_2 (e.g., image and text).
- Given a multimodal dataset $D = \{D^c, D^{m_1}, D^{m_2}\}$, the author denotes $D^c = \{x_i^{m_1}, x_i^{m_2}, y_i\}$ 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.
- To preserve the format of multimodal inputs, the author simply assigns 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

$$\widetilde{D}^{m_1} = \{x_j^{m_1}, \tilde{x}_j^{m_2}, y_j\}, \widetilde{D}^{m_2} = \{\tilde{x}_k^{m_1} x_k^{m_2}, y_k\}.$$

- Therefore, the multimodal data with missing modality can be reformed as $\widetilde{D} = \{D^c, \widetilde{D}^{m_1}, \widetilde{D}^{m_2}\}$.

Framework

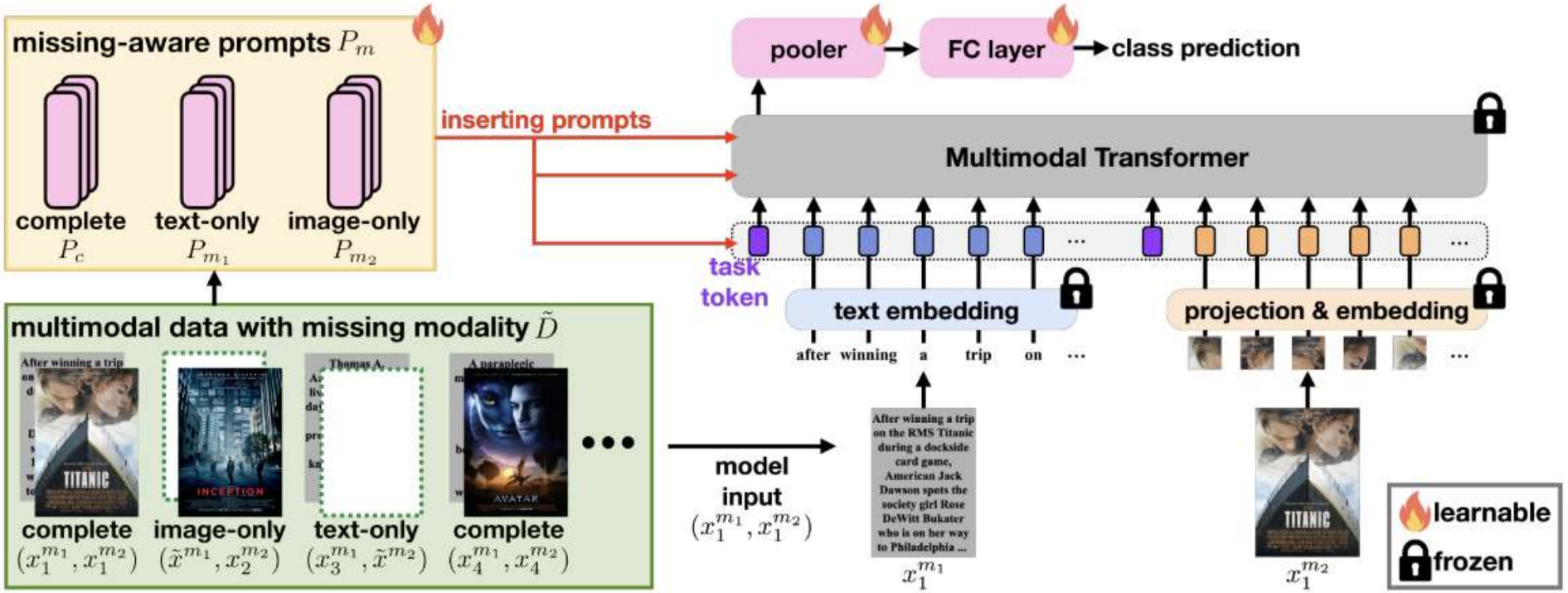


Figure 2. The overview of our proposed prompt-based multimodal framework. We first select the missing-aware prompts P_m according to the missing case (e.g., complete, text-only, image-only in vision-language tasks) of the multimodal inputs $(x_i^{m_1}, x_i^{m_2})$, in which the dummy inputs $\{\tilde{x}^{m_1}, \tilde{x}^{m_2}\}$ 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.

Framework

backbone:ViLT(untrainable)
In order to tackle the missing modality, the author proposes 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.

Given a pretrained multimodal transformer f_θ with N consecutive MSA layers, the author denotes the input embedding features of the i-th MSA layer as $h^i \in \mathbb{R}^{L\times d}, i = 1,2,\dots N$ with input length L and embedding dimension d.

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:

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

Framework

For model training, the author froze 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.

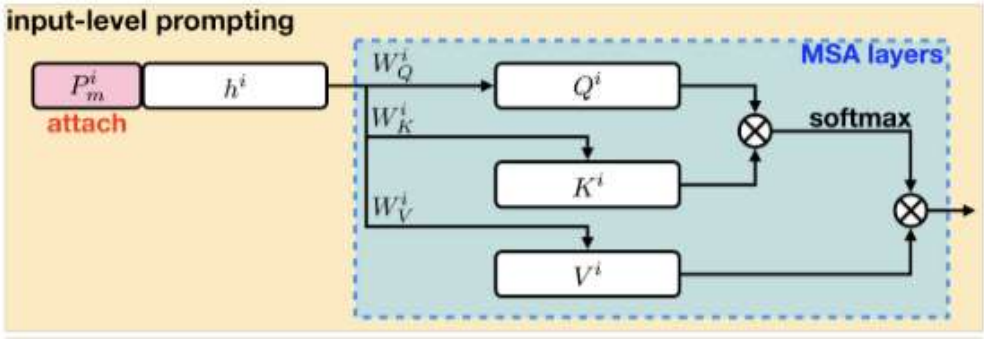
Moreover, the author denotes θ_p as the parameters of missing-aware prompts. The overall objective with trainable parameters is defined as:

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

$\left(x_i^{m_1}, x_i^{m_2}\right) \in \widetilde{D}$ 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.

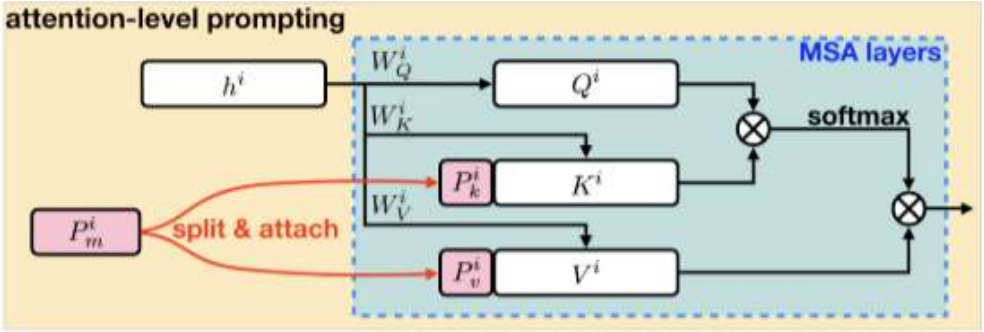
Method

Since the input modality may be missing, studying the proper configuration to attach prompts is of great importance. The author introduces two configurations of prompts: input-level prompting and attention-level prompting.



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

Method



the author 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.
the author denotes the query, key and value for the MSA layer as:

$$Q^i = h^i W_Q^i; K^i = h^i W_K^i; V^i = h^i W_V^i.$$

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

$$f_{prompt}^{attn}(p_m^i, h^i) = \text{ATTENTION}^i(p_m^i, h^i),$$
$$\text{ATTENTION}^i = softmax(\frac{Q^i[p_k^i, K^i]^T}{\sqrt{d}})[p_v^i; V^i].$$

Method

- Multi-layer prompting and locations where to attach prompts

$$P_m = \{p_m^i\}_{i=start}^{end} \in \mathbb{R}^{N_p \times L_p \times d}$$

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 is $N_p = (end - start + 1)$ the total number of layers with prompts.

Instead of attaching prompts to either whole layers or only the first layer, the author empirically finds that early half of layers is the best location starting from the first layer$\left(start = 0, end = \frac{N}{2} - 1\right)$ with $N_p = \frac{N}{2}$.

Experiment

- Datasets

  MM-IMDb ,UPMC Food-101, Hateful Memes

- Metrics

  different classification tasks

  For MM-IMDb , F1-Macro is adopted to measure the multi-label classification performance;

  For UPMC Food-101, the metric is the classification accuracy;

  For Hateful Memes , the metric is Area Underthe Receiver Operating Characteristic Curve (AUROC)

Experiment

Datasets	Missing rate η	Training		Testing		Baseline [13]	Attention-level prompts (Ours)	Input-level prompts (Ours)
		Image	Text	Image	Text			
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.

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

[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 su-pervision. In International Conference on Machine Learning(ICML), 2021. 1, 2, 4, 5, 6

[34]XinWang, 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

Experiment

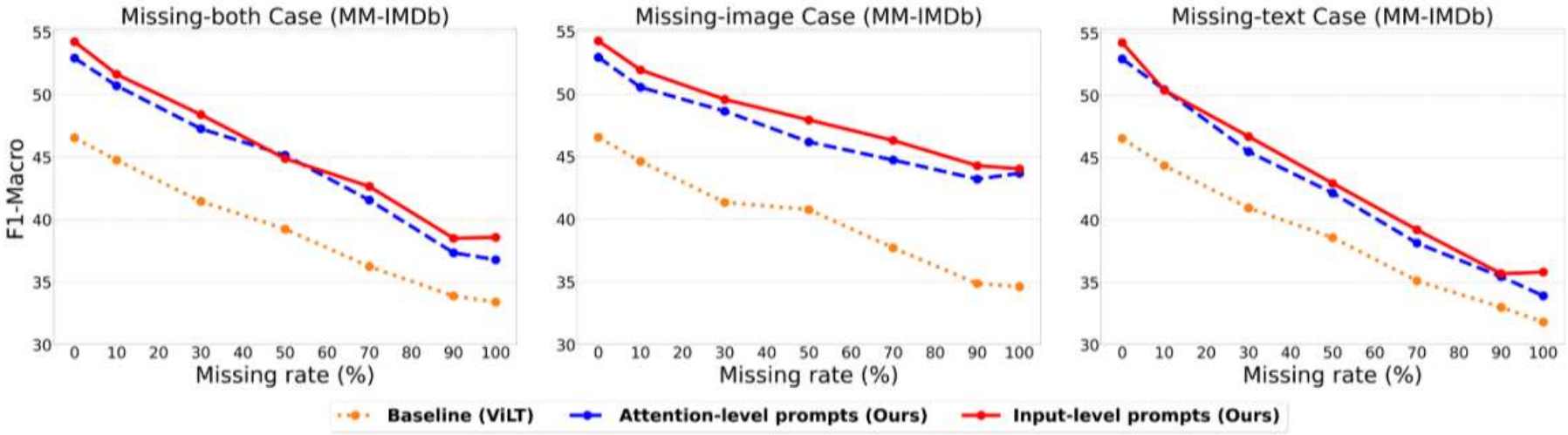
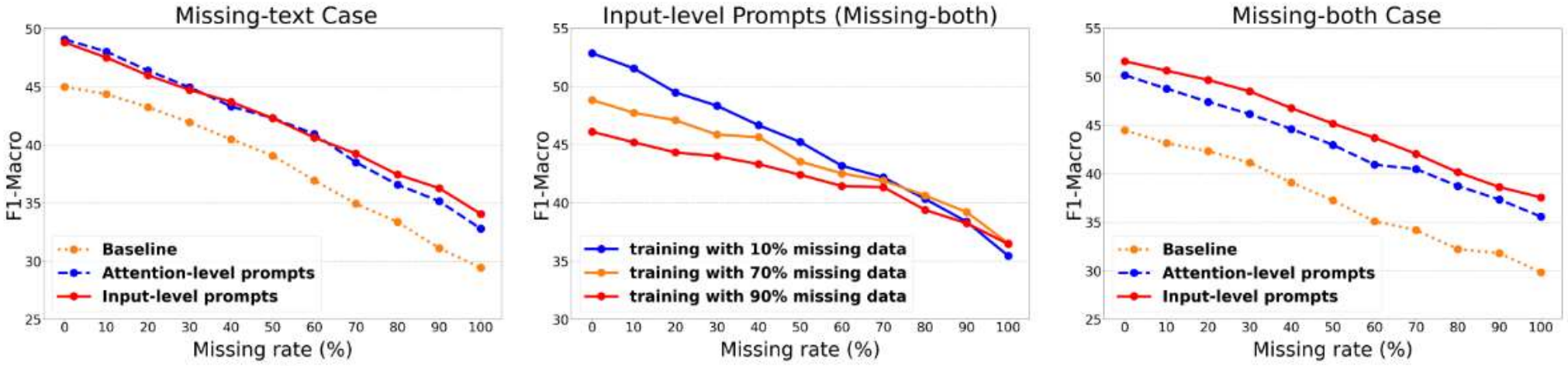


Figure 4. Quantitative results on the MM-IMDb dataset with different missing rates under different missing-modality scenarios. Each data point on the figure represents that training and testing are with the same $\eta\%$ missing rate.

Experiment



(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.

Experiment

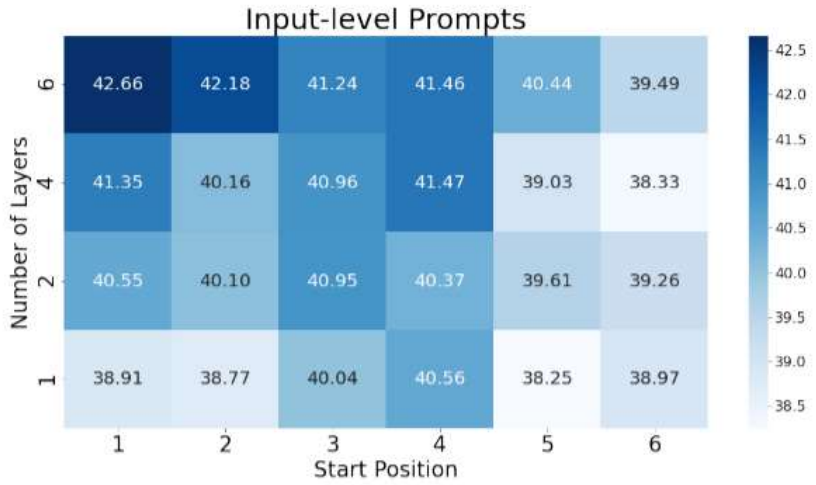


Figure 6. Ablation study on the location of prompting layers for input-level prompts.

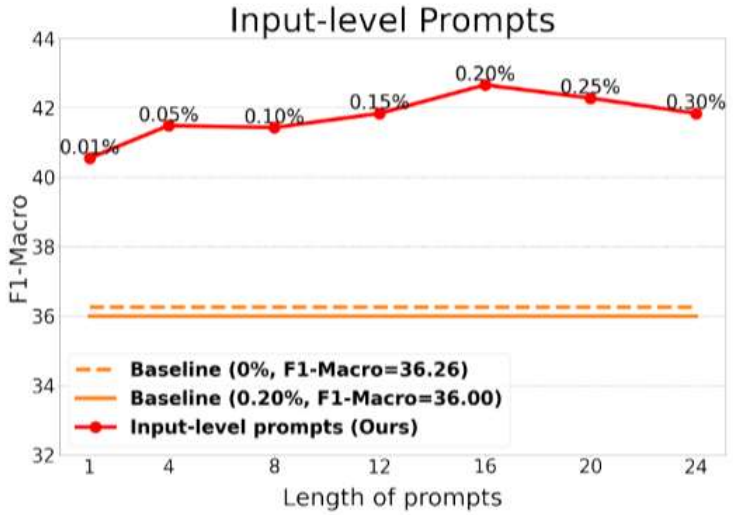


Figure 7. Ablation study on different length L_P of prompts for input-level prompts. The numbers above the red points are the proportion of parameters in prompts, compared to the entire model. We further conduct the new baseline with additional parameters with the same proportion (e.g., 0.2%) of the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

Thank you!