Multimodal Prompting with Missing Modalities for Visual Recognition
Supplementary Materials

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

1. More Results

1.1. Attention-level Prompts

We show ablation studies for attention-level prompts in Figure 1 and Figure 2, which analyze the effect of prompting layers and prompt length respectively. The results are similar to the study of input-level prompts as shown in Section 4.3 of the main paper. In summary, the earlier prompting layers and more prompting layers improve the performance more. In addition, even with fewer parameters (i.e., reducing the prompt length to 2), the performance is still competitive.

1.2. More Results on All Datasets

In Figure 3, we provide more quantitative results on different datasets (i.e., MM-IMDb [1], UPMC Food-101 [7], and Hateful Memes [3]) with different missing cases. The experiments are conducted by training on the specific missing rate $\eta\%$ and testing with the same rate. The trends are similar to the main results in Section 4.2 of the main paper. The proposed missing-aware prompts are able to tackle general missing modality cases without the need of finetuning the entire model. Moreover, the input-level prompts further show the favorable performance compared to the other two methods in most of the cases.

Particularly, we have discussed the sensitivity of input-level prompts to different datasets, which have slightly worse performance on Hateful Memes due to the long prompt length (i.e., prompt length is 16). Therefore, here we reduce the prompt length of input-level prompts to 4 for all experiments on Hateful Memes in Figure 3. As a result, the input-level prompting has competitive results with the attention-level prompting while consistently outperforming the baseline.

1.3. More Analysis

Robustness to different missing rates. We provide complete experiments of robustness to different missing rates which are mentioned in Section 4.3 of the main paper, as shown in Figure 4 and 5. With the similar trends as mentioned in the main paper, the results show that the input-level prompting is more robust to modality-incomplete data, while the performance of the other two methods depend on the training data composition.

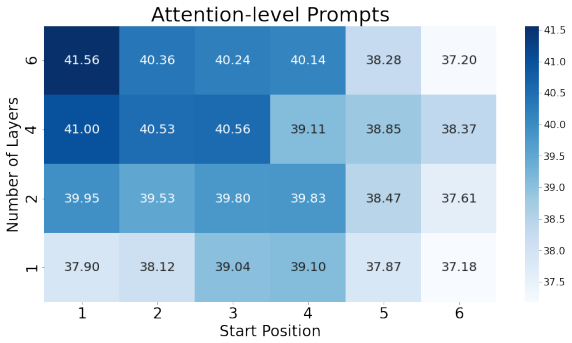

Figure 1. Ablation study on the location of prompting layers for attention-level prompts.

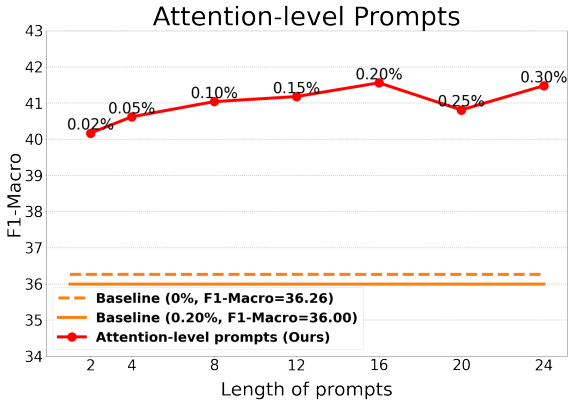

Figure 2. Ablation study on different length L_P of prompts for attention-level prompts. The numbers above the red points are the proportion of parameters in prompts, compared to the entire model. We further conduct the new baseline with additional parameters with the same proportion (e.g., 0.2%) of the prompt size, denoted as the orange solid line.

To be detailed, we have three observations in Figure 5:

1) The baseline is always worse than our prompting methods, and does not show obvious difference when train with missing rates 10% and 70%. With training on more modality-complete data (i.e., the 90% missing rate), it performs a bit better on modality-incomplete data but decreases the performance when testing on a low missing rate.

2) The performance of attention-level prompting depends on the training data composition. When training with more modality-incomplete data, the attention-level prompting performs well on a high missing rate but cannot reach the best performance on low missing rate.

3) The input-level prompts is more robust to the missing-incomplete data and reaches better performance with a low missing rate when training with more missing-complete data.

According to these observations, we prefer the input-level prompts more as its robustness to modality-incomplete data and better performance on different testing settings.

Results with complete training data. The entire results of training with modality-complete data are shown in Figure 6. We show that our method improves the baseline by a large margin when the training data is modality-complete. Moreover, the input-level prompts consistently improve the performance with different testing missing rates, showing that it is the more favorable prompting design on tackling the missing-incomplete data during testing.

Additional experiments on different multimodal tasks. To show the scalability of our proposed method, we evaluate our method on another audio-text sentiment analysis task following the transformer-based method [2], as shown in Table 1. We use the pre-trained model on the CMU-MOSEI [9] dataset and evaluate it on the MELD [6] dataset. Hence the scenario here becomes "adapting to different datasets with missing modalities". Even in such scenario, our missing-aware prompts still improve the model performance with missing modalities.

Methods	Training		Testing		Accuracy
	Audio	Text	Audio	Text	
Baseline	65%	65%	65%	65%	46.66
Attention-level prompts (ours)	65%	65%	65%	65%	**48.28**
Input-level prompts (ours)	65%	65%	65%	65%	47.47

Table 1. Quantitative results on MELD dataset.

More comparisons with prompt-based methods. Though many prompting techniques [5, 10] succeed in learning with multimodal downstream tasks without fine-tuning the model, they still suffer from the performance drop caused by missing modalities. In contrast, our proposed prompting method is modality-missing-aware, so it provides better instructions for tuning pre-trained backbone models in general modality missing scenarios. To better demonstrate the strength of our method, we perform additional experiments using *PromptFuse* and *BlindPrompt* proposed in [5] in the general setting. The results are shown in Table 2. We find that our method is more robust to missing modality with much higher F1-Macro performance.

Methods	Training		Testing		F1-Macro
	Image	Text	Image	Text	
PromptFuse [5]	100%	100%	65%	65%	31.21
	65%	65%	65%	65%	38.47
BlindPrompt [5]	100%	100%	65%	65%	33.16
	65%	65%	65%	65%	36.57
Input-level prompts (ours)	65%	65%	65%	65%	**42.66**

Table 2. Comparison with prompt-based baselines, PromptFuse and BlindPrompt, on the MM-IMDb dataset.

Efficiency and training speed. To demonstrate the efficiency of our proposed method, we evaluate the training speed on MM-IMDb including forward and backward respectively. As shown in Table 3, our method is 4.1× faster than standard finetuning during backward propagation, showing the efficiency of our missing-aware-prompts learning. Note that the forward time should be similar since the backbone model is the same.

Methods	forward time (ms)	backward time (ms)	speed up
Finetune	88.69	116.80	1x
Attention-level prompts	87.47	28.05	**4.16x**
Input-level prompts	90.80	28.10	4.15x

Table 3. The speed of forward and backward process during training on MM-IMDb dataset.

Limitations and future works. Although our modality-missing-aware prompting can largely increase the robustness of the tuned backbone models, it does not recover the missing information from the multimodal input. We expect the cross-modal generative modeling can help further boost the performance by generating missing information. Besides, when facing the scenario of having more and more modalities (i.e. increase in terms of number of modalities), there could be quadratic growth on number of prompts. To tackle this issue, we expect to adopt the prompt pool concept in the recent L2P [8] work, where the prompting mechanism can query from a fixed number of prompts in a designated pool to avoid the quadratic growth.

2. More Ablation Studies

We further investigate the effect of different prompt configurations with the same parameter size as well as the effect

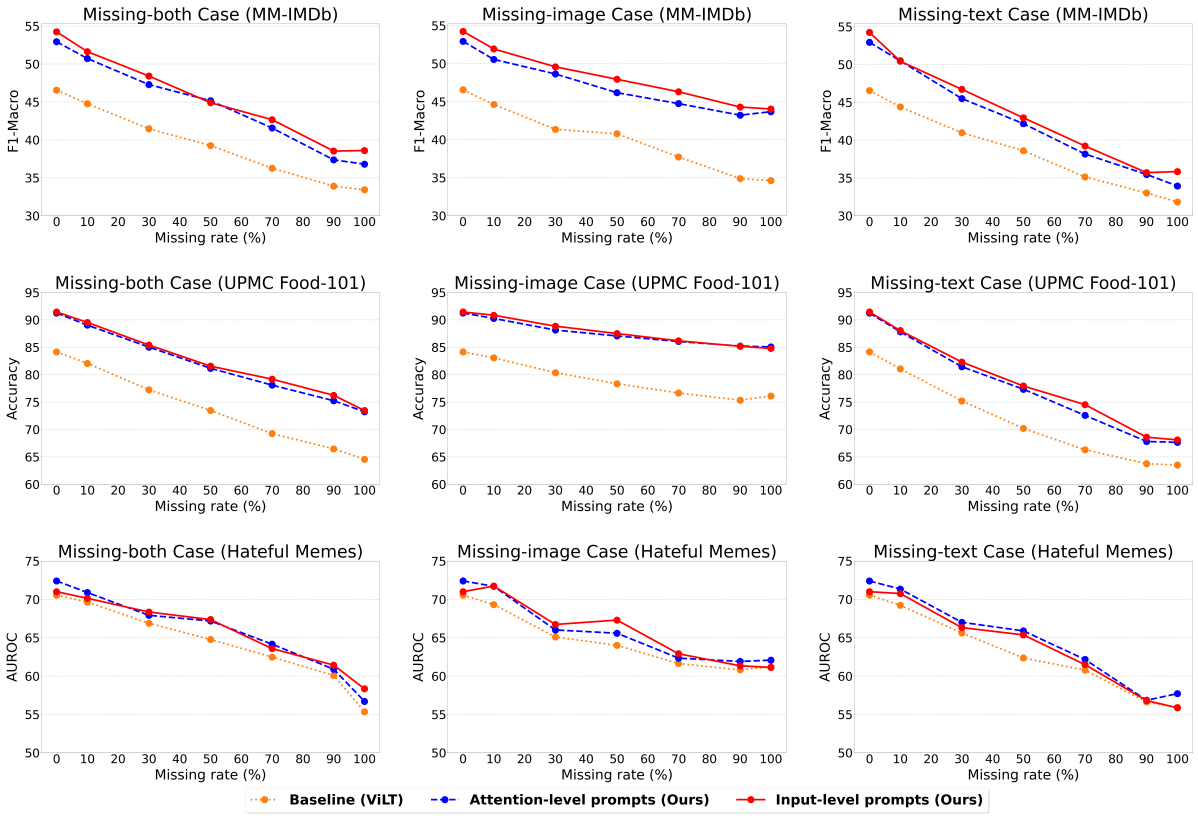

Figure 3. Quantitative results on the MM-IMDb, UPMC Food-101, Hateful Memes dataset with different missing rates under different missing-modality scenarios. Each data point on the figure represents that training and testing are with the same $\eta\%$ missing rate.

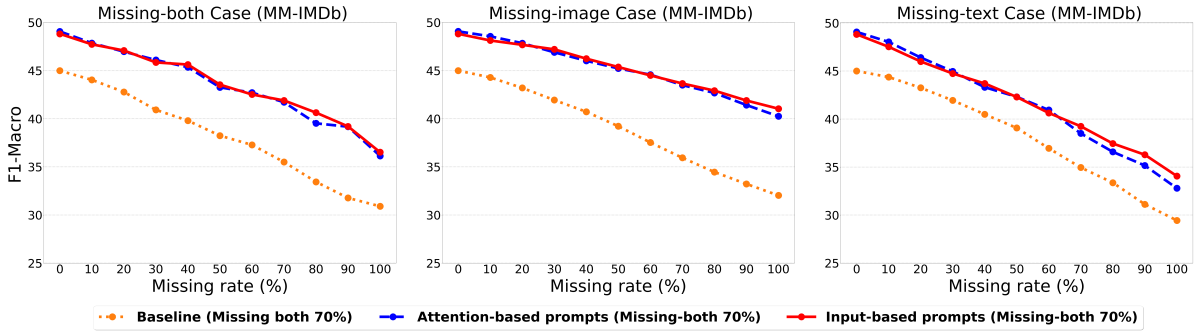

Figure 4. Ablation study on robustness to the testing missing rate in different scenarios on MM-IMDb. All models are trained on missing-both case with 70% missing rate, and evaluated on different cases with different missing rates.

of the alternative dummy inputs of missing images, which are shown in Table 4 and Table 5 respectively.

Fixed size of prompts. In the main paper, we have shown that the model with prompt length equal to 16 and attached prompts from the first layer to the sixth layer is the best configuration empirically. Here we compare the different configurations of prompts given a fixed size of parameters.

The results on MM-IMDb are shown in Table 4. Either models with fewer layers and a longer prompt length or models with more layers and shorter prompt length produce worse results. The missing-aware prompts attached to the early half of layers of the multimodal transformer with a suitable prompt length are the best choice for instructing the model performance.

Input for missing images In general, multimodal transformer allows the absence of any modalities via just masking out the missing inputs, thanks for its self-attention

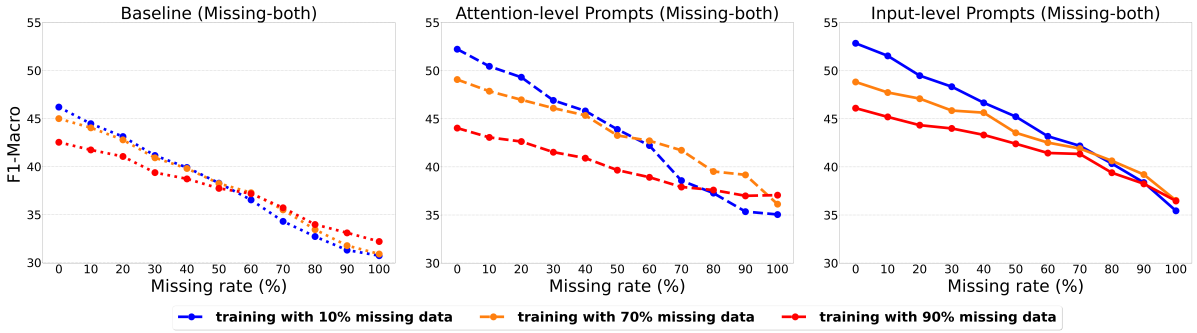

Figure 5. Ablation of different models trained on the missing-both cases with 10%, 70%, and 90% missing rates, which represent more modality-complete data, balanced data, and less modality-complete data, respectively. Evaluation is on missing-both case with different missing rates.

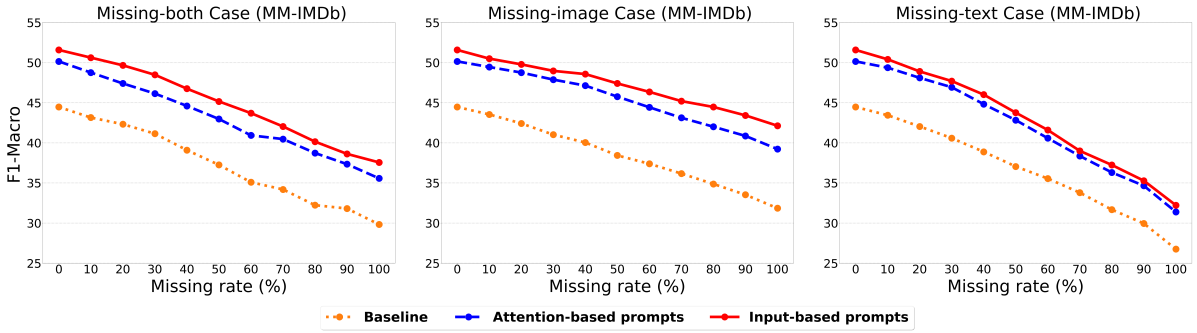

Figure 6. All models are trained with modality-complete data, where each data pair can be randomly assigned with different missing modality at different training epochs (i.e., text-only, image-only, and modality-complete) to account for possible missing modalities during testing. Evaluation is on missing-both case with different missing rates.

MM-IMDb [1]	Parameter size	# of layers N_p	Length L_p	F1-Macro
Attention-level Prompting	$N_p \times L_p \times D$	$2N$	$\frac{1}{2}L$	40.02
		$\frac{4}{3}N$	$\frac{3}{4}L$	40.85
		N	L	**41.56**
		$\frac{2}{3}N$	$\frac{3}{2}L$	40.88
		$\frac{1}{2}N$	$2L$	40.30
Input-level Prompting	$N_p \times L_p \times D$	$\frac{4}{3}N$	$\frac{3}{4}L$	41.21
		$2N$	$\frac{1}{2}L$	41.45
		N	L	**42.66**
		$\frac{2}{3}N$	$\frac{3}{2}L$	41.89
		$\frac{1}{2}N$	$2L$	41.08

Table 4. Different prompt configurations of prompts given the fixed size of parameters. The models are trained and tested on the missing-both case with missing rate $\eta\% = 70\%$ on MM-IMDb. The default value of N, L, D is 6, 16, 768, respectively. **Bold** numbers indicate the best performance.

mechanism to generate a holistic representation of all modalities. Apart from masking, we also can generate a dummy sample to represent the missing modalities. Here we compare different ways to deal with the missing-image case: masked image, all-one image, all-zero image, fixed random image, instance-wise random images and iteration-wise random images.

- **Masked image**: Mask out image tokens when the image is missing.
- **All-one image**: A dummy image tensor with all one values.
- **All-zero image**: A dummy image tensor with all zero values.
- **Fixed random image**: A pre-defined dummy image tensor with random samples from normal distribution $N(0, 1)$ in advance.
- **Instance-wise random images**: Assign a specific random alternative image tensor sampled from $N(0, 1)$ for each data pair with missing image.
- **Iteration-wise random images**: Randomly generate alternative images for pairs with missing-image cases in each iteration during training.

The results are shown in Table 5. We find that the "masked image" is the best way for the baseline model, since the model can ignore the attention on missing values by masks and eliminate the effect of missing modality. In contrast, our prompt-based models perform better with "fixed random image" and "all-one image". This shows that with instruction of prompts, the model can process the dummy images effectively. We find that "all-one image" works well in most cases and thereby set it as the default setting in our experiments.

MM-IMDb [1]	missing image inputs	F1-Macro
Baseline (ViLT [4])	Masked image	**36.28**
	All-one image	<u>36.26</u>
	All-zero image	30.73
	Fixed random image	34.95
	Instance wise random images	36.18
	Iteration-wise random images	35.35
Attention-level Prompting	Masked image	40.45
	All-one image	<u>41.56</u>
	All-zero image	40.06
	Fixed random image	**41.84**
	Instance wise random images	41.35
	Iteration-wise random images	41.05
Input-level Prompting	Masked image	42.39
	All-one image	**42.66**
	All-zero image	41.42
	Fixed random image	42.51
	Instance wise random images	42.33
	Iteration-wise random images	<u>42.59</u>

Table 5. Ablation study on the selection of alternative inputs for missing images. **Masked image:** mask out image tokens when the image is missing. **All-one (all-zero) image:** a dummy image tensor with all one (zero) values. **Fixed random image:** a pre-defined alternative image input randomly sampled from normal distribution. **Instance-wise random images:** assign random alternative image input for each data pair. **Iteration-wise random images:** randomly generate alternative images in each iteration during training. **Bold** numbers indicate the best performance and <u>underlined</u> numbers are the second best one.

Final output feature selection. By default, we follow ViLT, which is also our pre-trained multimodal backbone, to use the text-related task token as the final output feature. In addition, we have evaluated different output features as outputs. The text-related and image-related tokens are the pre-trained tokens and kept frozen during training, while the independent task token is another learnable token added in front of the text-related token (*i.e.*, the first token of entire input sequence) and kept updated during training. As shown in Table 6, we find the default setting (*i.e.*, using the text-related task token as in ViLT) works the best.

Methods	Output features (task token)		
	Text-related	Image-related	Independent
Baseline	36.26	29.57	34.22
Attention-level prompts	41.56	37.93	39.39
Input-level prompts	42.66	41.23	41.74

Table 6. The ablation study of the choice on final output features.

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. 1, 4, 5

[2] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv preprint arXiv:2006.15955*, 2020. 2

[3] 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. 1

[4] 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. 5

[5] 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. 2

[6] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018. 2

[7] 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. 1

[8] 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. 2

[9] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018. 2

[10] 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. 2