Supplementary: Missing Modality Prediction for Unpaired Multimodal Learning via Joint Embedding of Unimodal Models

A Implementation Details

We conduct overall experiments using a single RTX 3090 GPU. Similar to [4], we employ DeiT III [7] as an image encoder and SimCSE [3] as a text encoder throughout the overall experiment, which model parameters are initialized using the pretrained weights. The structure of the feature predictor is modified from [2], which consists of two fully-connected layers with layer normalization layer [1] and activation function, and a third linear layer. The dimensions of all three layers are set to equal to the output dimension of the encoders. The length of learnable read-only prompts for feature prediction is set to 6 in MM-IMDb, 20 in UPMC Food-101, and 2 in Hateful Memes in both training settings. We conduct all experiments with batch size 12 of 20 epochs, using the AdamW [6] optimizer with a weight decay of 5×10^{-2}. We initiate the warm-up steps for the learning rate at 0, with a base learning rate set to 1×10^{-2}. The warm-up phase linearly progresses from 0 for the first 10% of the total training steps before decay. The variance, invariance, and covariance coefficients λ, μ and ν for VICReg loss are set to 50, 50, and 1.

For the ablation study on section **??**, we add 36 learnable tokens to the input data specifically for the target task in prefix tuning, separate from utilizing read-only prompts for feature prediction. For adapter tuning, the adapter is inserted before each layer normalization layer in transformer-based pretrained models. It consists of a weight matrix that downsamples the input, followed by an activation function, and another weight matrix that restores the input to its original dimension. Finally, it includes a residual connection. We set the reduction factor to 4. For a fair comparison that considers trainable parameters, adding an adapter is applied to the first and last layers, and separately to all layers for further analysis.

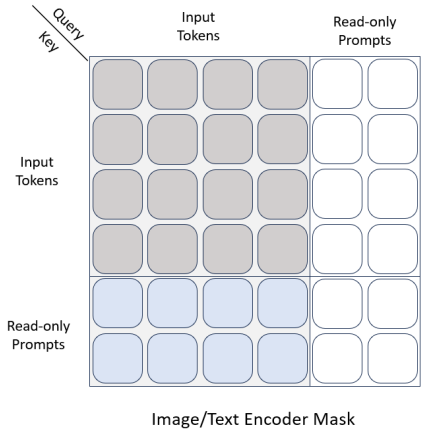

Fig. 1: The attention mask we use in Read only prompts. The colored boxes indicate areas where attention operation occurs, while the white boxes indicate masked regions.

B Read Only Prompts Attention Mask

As illustrated in Fig. 1, we applied an attention masking mechanism for our encoder's architecture, following the approach outlined in [5]. This structure is essential to maintain the representations of the input tokens for the target task while training the learnable tokens specialized for feature prediction.

C Additional Results

C.1 Results on Image Missing Case

We provide a result of an image missing case in the Fig 2 under the complete training with the Food-101, which is the most susceptible among the three datasets. Due to the dominance of the text modality, the impact of the image modality on target task is minimal. However, the unimodal baseline suffers in severely image-missing cases. Conversely, our method, which demonstrates superior performance in the text missing case, also shows robustness to severely image missing cases

C.2 Results on Different Missing Ratio

As illustrated in Fig. 3, we explored varying missing rates from the fixed 70% on [both missing training/testing] setting. *Our method with feature predictor trained on 70% missing data (i.e., only 30% paired) outperforms the unimodal baseline trained with fully-paired data.* When the missing rate is exceptionally high (i.e., 90% missing), feature predictors struggle to learn effectively from the limited

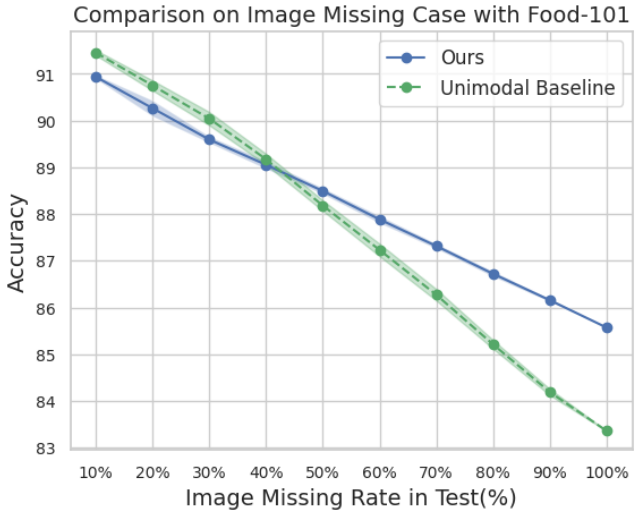

Fig. 2: Additional results on image missing case under complete training setting. Dotted lines denote unimodal baseline, whereas solid lines represent ours.

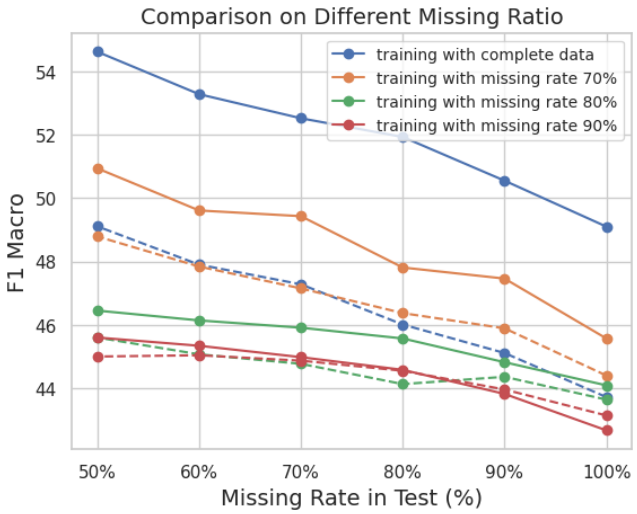

Fig. 3: Additional results on different missing ratio under missing training setting. Dotted lines denote unimodal baseline, whereas solid lines represent ours

paired data available. Nonetheless, even with just 20% paired data, our method surpasses the unimodal baseline.

C.3 Full Results of the Ablation Study on prompts-based feature prediction

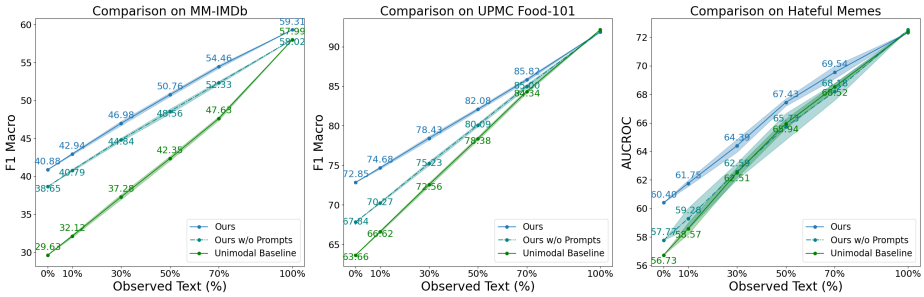

Fig. 4: Additional results on ablation study about the effect of prompts-based feature prediction under a complete training setting. All experiments were conducted with 100% text and 100% image data and evaluated based on the text missing rate. For a fair comparison, we evaluate by averaging performances across five different random seeds. Ours w/o prompts leverage CLS token for feature prediction.

C.4 Full Results on Missing Training Setting

Table 1: Full comparison results under a training setting with a 70% missing rate. Evaluations were conducted by averaging the results from five different seeds across various modality-missing scenarios.

Dataset	Training Image	Training Text	Testing Image	Testing Text	ViLT	Lee et al.	Unimodal Baseline	Ours w/o Prompts	Ours
MM-IMDB (F1-Macro)	100%	30%	100%	30%	32.78±0.29	37.72±0.21	39.47±0.17	40.01±0.19	**43.21**±0.19
			30%	100%	26.58±0.29	21.68±0.32	42.12±0.07	49.72±0.17	**54.67**±0.14
			65%	65%	30.55±0.17	30.80±0.59	40.99±0.06	45.02±0.12	**49.09**±0.12
			AVG		29.97	30.07	40.86	44.91	**48.99**
	30%	100%	100%	30%	30.25±0.30	24.93±0.12	29.85±0.13	39.60±0.25	**43.07**±0.31
			30%	100%	37.97±0.21	47.10±0.28	54.37±0.16	50.94±0.26	**56.03**±0.14
			65%	65%	34.45±0.44	36.76±0.27	37.61±0.24	45.83±0.15	**49.81**±0.16
			AVG		34.22	36.26	40.61	45.46	**49.64**
	65%	65%	100%	30%	35.80±0.29	39.04±0.15	40.60±0.26	41.61±0.11	**42.46**±0.24
			30%	100%	36.65±0.16	42.68±0.10	53.19±0.26	52.05±0.14	**55.26**±0.22
			65%	65%	36.66±0.27	41.33±0.48	47.34±0.24	47.29±0.11	**49.24**±0.50
			AVG		36.37	41.02	47.04	46.98	**48.99**
Food-101 (Accuracy)	100%	30%	100%	30%	66.02±0.18	73.51±0.24	78.60±0.11	78.50±0.18	**78.81**±0.13
			30%	100%	42.89±0.22	27.55±0.06	82.50±0.14	86.03±0.14	**86.90**±0.10
			65%	65%	54.49±0.12	50.59±0.16	80.37±0.22	82.22±0.08	**82.35**±0.14
			AVG		54.47	50.55	80.49	82.25	**82.69**
	30%	100%	100%	30%	42.80±0.31	29.61±0.16	66.50±0.20	73.97±0.16	**75.41**±0.22
			30%	100%	76.54±0.12	86.45±0.11	**87.44**±0.37	87.17±0.11	87.11±0.11
			65%	65%	59.54±0.19	57.97±0.15	76.98±0.10	80.49±0.17	**81.22**±0.18
			AVG		59.63	58.01	76.97	80.55	**81.25**
	65%	65%	100%	30%	64.40±0.28	71.26±0.30	76.51±0.20	77.67±0.16	**78.13**±0.15
			30%	100%	73.60±0.13	85.71±0.15	87.21±0.14	86.79±0.12	**87.35**±0.12
			65%	65%	68.96±0.27	78.43±0.16	81.82±0.14	82.18±0.12	**82.67**±0.16
			AVG		68.99	78.47	81.85	82.21	**82.72**
Hateful Memes (AUROC)	100%	30%	100%	30%	60.77±0.72	58.54±1.27	59.77±0.63	57.65±0.50	**61.39**±0.82
			30%	100%	61.84±0.42	55.67±1.30	66.41±0.28	67.36±0.25	**66.50**±0.36
			65%	65%	61.57±0.93	57.10±1.73	61.86±2.04	62.58±1.36	**63.35**±1.18
			AVG		61.39	57.10	62.68	62.53	**63.75**
	30%	100%	100%	30%	56.19±0.86	58.43±0.37	58.98±0.57	59.04±1.31	**62.06**±0.38
			30%	100%	62.78±0.74	65.54±0.61	68.52±0.37	**70.53**±0.10	68.97±0.30
			65%	65%	59.29±1.02	62.36±0.35	63.58±0.16	64.78±0.98	**65.12**±0.88
			AVG		59.42	62.11	63.69	64.78	**65.38**
	65%	65%	100%	30%	59.42±0.92	60.05±1.40	60.93±0.46	60.39±0.27	**61.39**±0.58
			30%	100%	63.13±0.45	61.88±1.07	**69.19**±0.28	66.30±0.47	68.24±0.28
			65%	65%	61.32±1.21	61.52±1.11	64.47±0.63	62.90±0.47	**65.17**±0.79
			AVG		61.29	61.15	64.86	63.20	**64.93**

C.5 Feature Prediction Loss Coefficients

Table 2: Coefficient for feature prediction loss under a complete training setting.

λ	μ	ν	F1-Macro
5	5	1	36.11
15	15	1	39.19
25	25	1	38.42
50	50	1	**40.88**

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021)
3. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821 (2021)
4. Khan, Z., Fu, Y.: Contrastive alignment of vision to language through parameter-efficient transfer learning. arXiv preprint arXiv:2303.11866 (2023)
5. Lee, D., Song, S., Suh, J., Choi, J., Lee, S., Kim, H.J.: Read-only prompt optimization for vision-language few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1401–1411 (2023)
6. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
7. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. In: European Conference on Computer Vision. pp. 516–533. Springer (2022)