Supplementary Material for Multi-View Slot Attention Using Paraphrased Texts for Face Anti-Spoofing

Jeongmin Yu[1*], Susang Kim[1,2*], Kisu Lee[1], Taekyoung Kwon[1], Won-Yong Shin[1], Ha Young Kim[1†]
[1]Yonsei University [2]POSCO DX
{jeongminyu, healess, kisu0928, taekyoung, wy.shin, hayoung.kim}@yonsei.ac.kr

A. Multi-View Approach on Fixed Prompt

To compare learnable prompts with non-trainable fixed prompts, we conduct an ablation study following the same procedure as in Sec. 4.3.2 and Sec. 4.3.3, employing the fixed prompt pairs (Table A.1) from FLIP [6]. As shown in Table A.2, some fixed prompt pairs achieve comparable performance to learnable prompts in a single-view setting. However, Table A.3 indicates that fixed prompt pairs underperform in the multi-view setting compared to using learnable context texts $[\mathbf{V}]$. We infer that the lower synergy in the multi-view setting occurs because heuristically determined fixed prompt pairs cannot guarantee the optimal context for the model.

Pair No.	Positive prompts	Negative prompts
P1	This is an example of a real face	This is an example of a spoof face
P2	This is a bonafide face	This is an example of an attack face
P3	This is a real face	This is not a real face
P4	This is how a real face looks like	This is how a spoof face looks like
P5	A photo of a real face	A photo of a spoof face
P6	This is not a spoof face	A printout shown to be a spoof face

Table A.1. Fixed prompt pair settings. Each prompt pair is used in FLIP [6].

Single-view	OCI → M	OMI → C	OCM → I	ICM → O	avg.
	HTER(%)↓	HTER(%)↓	HTER(%)↓	HTER(%)↓	HTER(%)↓
P1	3.92	3.90	7.47	3.94	4.81
P2	5.00	4.71	3.54	5.85	4.78
P3	3.96	**2.84**	4.96	4.08	**3.96**
P4	4.08	4.30	4.28	**3.29**	3.99
P5	**3.42**	4.59	4.66	4.58	4.31
P6	4.33	5.99	**3.06**	5.54	4.73

Table A.2. Comparison of HTER among single-view text pairs in Protocol 1.

B. Ablation Study on MTPA Coefficient

For the additional loss $\mathcal{L}_{\text{MTPA}}$, we empirically determine its optimal weight λ in the total loss. To evaluate its impact on performance, we use the total loss as follows: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{MTPA}}$. As shown in Fig. B.1, we compare the results when the scale of $\mathcal{L}_{\text{MTPA}}$ decreases ($\lambda < 1.0$), remains the same ($\lambda = 1.0$), or increases ($\lambda > 1.0$). The experiments indicate that the setting with $\lambda = 1.0$ achieves the best performance among all settings, underscoring the importance of balancing alignment and classification. Consequently, we set the coefficient of $\mathcal{L}_{\text{MTPA}}$ to 1.0 in the main paper.

Number	HTER↓	AUC↑	TPR@FPR=1%↑
M=1	3.96	99.03	84.89
M=2	3.36	99.33	86.23
M=3	3.11	99.19	85.48
M=4	2.90	**99.46**	**89.55**
M=5	2.81	99.30	85.69
M=6	**2.67**	99.35	85.30

Table A.3. Ablation study on the number of multi-views for fixed prompts, with each result (%) representing the average across all scenarios in Protocol 1.

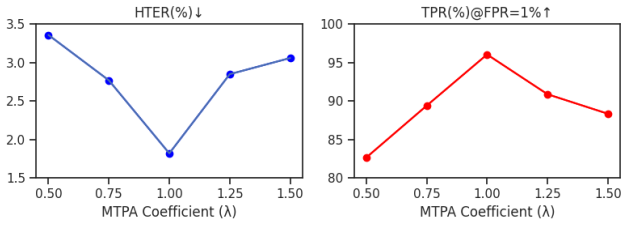

Figure B.1. Performance of HTER and TPR(%)@FPR=1% according to the MTPA coefficient, with each result representing the average across all scenarios in Protocol 1.

C. Additional Visualization Results of MVS

Additional visualization results of MVS for positive and negative samples are provided in Fig. C.1.

D. Effectiveness of soft-masking in MTPA

To evaluate the impact of soft-masking in MTPA, we conduct an ablation study, comparing the performance

*Equal contribution
†Corresponding author

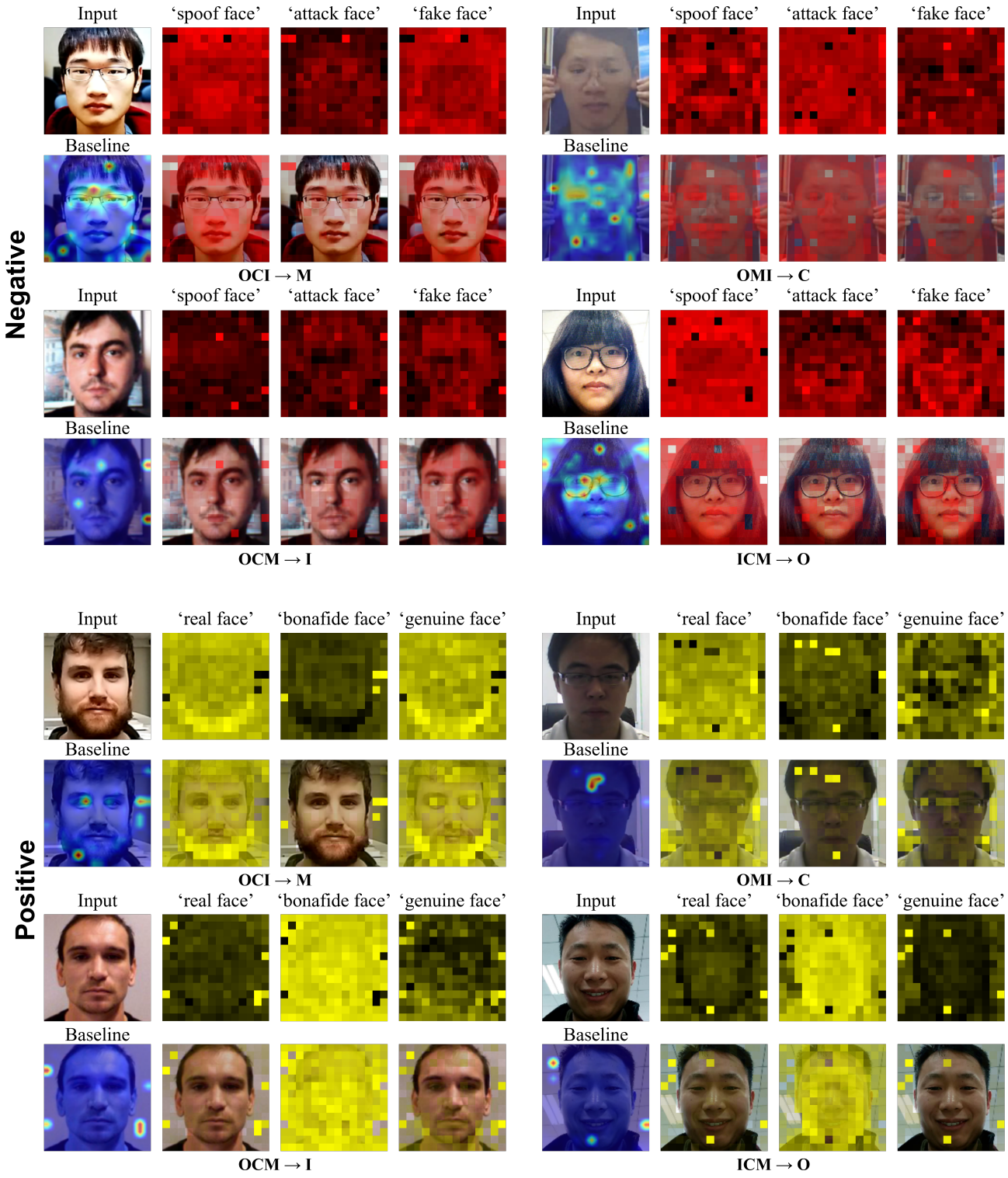

Figure C.1. **Additional visualization results for each view on real and spoof images across all sub-protocols in Protocol 1.** We visualize multi-view attention scores in the first stage of MVS. Baseline (Sec. 4.3.1) indicates the attention map [2], following the visualization method used in previous studies [3, 4, 6]. The first row for each cross-domain dataset shows the MVS visualization corresponding to the tree prompts, while the second row displays the MVS attention maps overlaid on the input images.

of MTPA with and without soft-masking. Without soft-masking, it is equivalent to attaching an auxiliary classifier to the model while ignoring the similarity between patches and text. Consequently, as shown in Table D.1, without soft-masking shows lower performance compared to with soft-masking, which aligns important patches for spoofing prediction based on patch-text similarity. These results indicate that similarity-based soft-masking in MTPA offers more effective alignment guidance to the model.

	HTER↓	AUC↑	TPR@FPR=1%↑
MTPA (w/o soft-masking)	3.90	99.13	81.79
MTPA (w/ soft-masking)	**1.82**	**99.70**	**96.06**

Table D.1. Ablation study on the effectiveness of soft-masking in MTPA, with each result (%) representing the average across all scenarios in Protocol 1.

E. Effectiveness of Prompt Subsets in MVS

Subset of prompts	OCI → M	OMI → C	OCM → I	ICM → O	avg.
	HTER(%)↓	HTER(%)↓	HTER(%)↓	HTER(%)↓	HTER(%)↓
Lowest	1.83	**0.45**	2.72	**2.56**	1.89
Highest	**1.71**	0.75	**1.99**	2.82	**1.82**

Table E.1. Ablation study on the effectiveness of different subsets of prompts in multi-view setting ($M = 3$) in Protocol 1.

We conduct an ablation study to evaluate the effectiveness of different subsets of prompts in the multi-view setting ($M = 3$). We selected prompts based on single-view performance in Sec. 4.3.2 including those with the **Highest** and **Lowest** performance. Due to computational resources, we could not evaluate all possible combinations of prompts (${}_5C_3 \times 4$). Nevertheless, as shown in Table E.1, our method still yields consistent performance improvements (HTER%) even when combining three prompts with the lowest single-view performance.

F. Multi-View Text Generation

To generate multi-view texts, we use this sentence to generate paraphrased texts: {**Paraphrase the sentences "real face" and "spoof face" into sentences of the form "xxxx face" with the same semantic meaning for anti-spoofing.**} In this paper, we manually assessed the quality of the texts generated by ChatGPT [1] and excluded those of insufficient quality.

G. Semantic Diversity in Class Texts

This paper aims to extract diverse features for generalization by using various paraphrased texts rather than prompt tuning in previous works [3, 5]. As shown in the figure G.1,

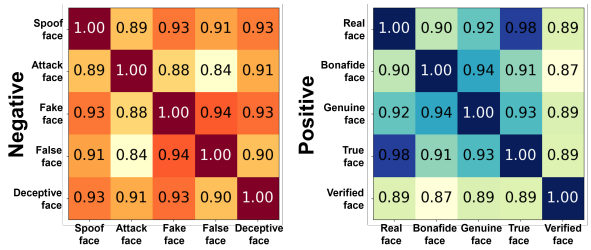

Figure G.1. Similarity for negative and positive texts. The heatmap shows feature similarity among five similar texts within each group. Darker colors indicate higher similarity.

which shows the similarity between CLIP text embeddings for each class text, even short class texts can have varied embeddings. Based on these results, instead of including diverse attributes in the prompts, we use simple class texts to avoid the influence of other attributes.

H. Details of MVS Ablation Study

This section provides detailed settings for two alternative methods in Table 7 of Sec. 4.3.5 and explains the reasoning behind their selection.

Similarity. To compare our MVS with the text-image information fusion method used in FLIP [6], we select a similarity-based approach as a baseline. Specifically, this method makes predictions by computing the cosine similarity between the mean of global-aware patch embeddings passed through a projection layer and the mean text embeddings of each class, without using a classification head.

Cross-attention. Unlike slot attention, the general attention mechanism [7] combines values through weighted summation. Therefore, using patch embeddings as queries and text embeddings as keys and values intermixes the semantic meanings of positive and negative texts, which results in significant performance degradation (average HTER of 14.67 in Protocol 1). Consequently, to effectively aggregate patch information, text embeddings as queries and patch embeddings as keys and values are more suitable for cross-attention. Thus, we adopt this setting to evaluate cross-attention. The cross-attention setting, similar to MVS, computes the mean of the cross-attention module's output, projects it, and passes it through a classification head to produce the prediction results.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder

transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 2

[3] Xueli Hu, Huan Liu, Haocheng Yuan, Zhiyang Fu, Yizhi Luo, Ning Zhang, Hang Zou, Gan Jianwen, and Yuan Zhang. Fine-grained prompt learning for face anti-spoofing. In *ACM Multimedia*, 2024. 2, 3

[4] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 222–232, 2024. 2

[5] Si-Qi Liu, Qirui Wang, and Pong C Yuen. Bottom-up domain prompt tuning for generalized face anti-spoofing. In *European Conference on Computer Vision*, 2024. 3

[6] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19685–19696, 2023. 1, 2, 3

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3