SUPPLEMENTAL MATERIALS FOR XSA-MAD: CROSS-MODAL SEMANTIC ALIGNMENT FOR MORPHING ATTACK DETECTION

Jie Jin[1,2]*, Mahiro Tokumasu*[2]*, Yu Makino*[2]*, Masakatsu Nishigaki*[1,2]*, Tetsushi Ohki*[1,2,3]

Graduate School of Science and Technology[1] Shizuoka University[2], RIKEN AIP[3]

1. REPRESENTATIVE STRUCTURED PROMPTS

Table 1 presents representative examples of the structured prompts used in this study. Due to space limitations, the complete prompt list is not included; however, all prompts share the same structural template and differ only in their attribute-specific textual content.

Each structured prompt consists of a fixed *Base Context* followed by four attribute-specific descriptions corresponding to *Identity*, *Geometry*, *Texture*, and *Consistency*. The listed examples, drawn from the prompt pool used in Section 2.2, illustrate typical linguistic patterns and semantic variations across bona-fide and morph classes, and were generated using the same deterministic procedure as the full prompt set, without manual post-editing or class-specific refinement.

2. TEXTUAL SPACE RESTRUCTURING

The structural transformation of the textual semantic space during Stage 1 optimization is visualized in Fig. 1. In the default CLIP configuration (Fig. 1a), bona-fide and morph classes exhibit significant entanglement without a clear discriminative boundary. Attribute anchors are closely clustered, failing to provide class-specific differentiation. Upon Stage 1 optimization, the space undergoes a fundamental restructuring into two distinct, well-isolated clusters (Fig. 1b). Prompt features align tightly around attribute anchors, establishing the stable decision boundaries necessary for robust detection.

3. IMPACT OF ATTRIBUTE ABLATION

3.1. Quantitative Performance Degradation

Table 2 summarizes the detection performance when one of the four semantic attributes—Identity (I), Geometry (G), Texture (T), or Consistency (C)—is removed from the structured prompts. The overall results show that the proposed XSA-MAD configuration achieves an EER of 6.13%. In comparison, the variants utilizing three attributes show higher overall error rates across the evaluated datasets.

This work was supported in part by JSPS KAKENHI JP23H00463, JP23K28085, and JST Moonshot R&D Grant Number JPMJMS2215.

Specifically, configurations without the Identity (w/o I) or Texture (w/o T) attributes result in the highest overall EERs, recorded at 7.27% and 7.26% respectively. Similarly, configurations without Geometry (w/o G) or Consistency (w/o C) also show a degradation in performance compared to the full model, yielding overall EERs of 6.89% and 6.99%. Across these different ablation settings, the increase in the overall EER ranging from 0.76 to 1.14 percentage points is consistently observed when any one attribute is removed.

These quantitative results demonstrate that, among all tested configurations in Table 2, the set comprising all four semantic attributes achieves the lowest overall error rate.

3.2. Qualitative and Structural Analysis

The structural visualizations in Fig. 1 characterize how the textual semantic space changes under different attribute ablation settings. With the full XSA-MAD configuration, the space maintains a clear structure, achieving a robust inter-class separation of 2.12 and intra-class coherence of 0.43 for both bona-fide and morph clusters. In contrast, removing any single semantic axis consistently degrades this structure, leading to a less stable and less discriminative embedding space.

Specifically, as shown in Fig. 1c, removing the Identity attribute reduces the inter-class separation to 1.79 while yielding intra-class variances of 0.44 and 0.58. The absence Texture, depicted in Fig. 1e, induces the most severe contraction of the separation distance to 1.46 with coherence values of 0.59 and 0.47 for bona-fide and morph features, respectively. When Geometry is excluded as visualized in Fig. 1d, the UMAP projection shows that the bona-fide and morph prompt clusters exhibit poor cohesion. However, this results in a visually diminished inter-class distance that aligns with the drop in angular separation to 1.58. The intra-class coherence of 0.65 for bona-fide and 0.48 for morph remains comparable to the levels of degradation observed in other ablation settings. The exclusion of Consistency as depicted in Fig. 1f lowers the inter-class separation to 1.75 and results in a bona-fide intra-class variance of 0.62 alongside 0.50 for morph.

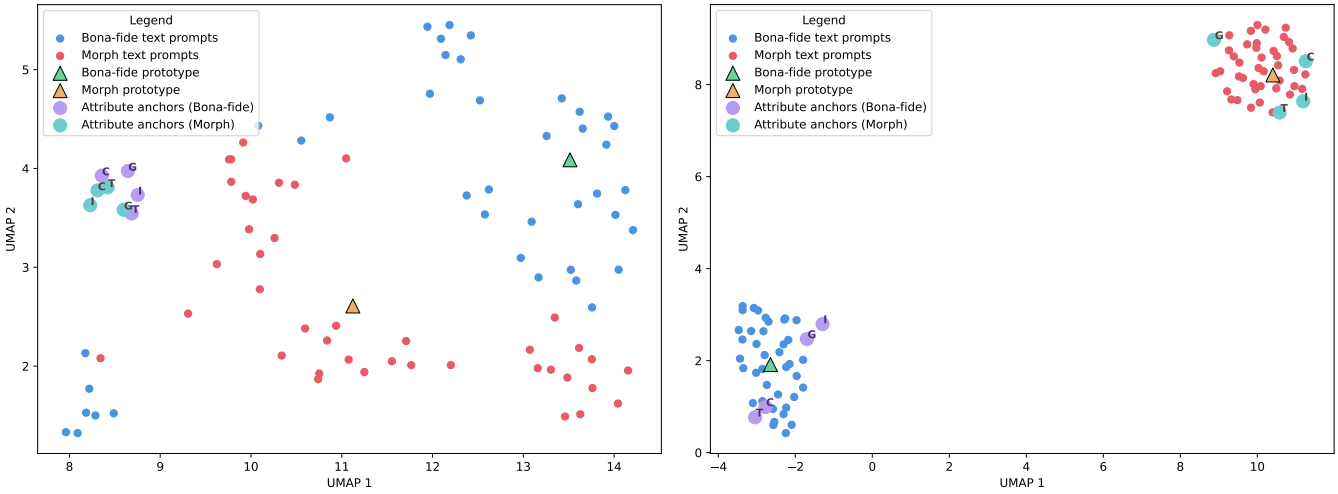
Collectively, these observations regarding visual restructuring and spatial metrics demonstrate that the presence of all four attributes is essential for the stable construction of class prototypes.

Table 1: Representative examples of structured prompts.Each prompt consists of a fixed Base Context followed by four attribute-specific sentences,each explicitly introduced by its attribute name(Identity, Geometry,Texture, and Consistency).

ID	Structured Prompt (Base Context with Identity, Geometry, Texture, and Consistency)
B01	A portrait of a human face. Identity: A single consistent human identity belonging to one real individual. Geometry: Biologically coherent facial proportions with no distortion. Texture: Natural skin texture showing pores and subtle imperfections. Consistency: All biometric signals support one unified subject.
B02	A natural facial portrait. Identity: One unique person represented authentically. Geometry: Stable landmark alignment that reflects a real subject. Texture: Fine details preserved without smoothing. Consistency: The entire image reflects a single identity origin.
B03	A standard passport-style image. Identity: Captured from one real person. Geometry: Symmetric facial arrangement aligned naturally. Texture: Authentic micro-texture without artificial blending. Consistency: No visual cues indicate identity mixing.
B04	A frontal biometric capture. Identity: Clear representation of one human identity. Geometry: Normal human facial dimensions and alignment. Texture: Naturally varied textures typical of live skin. Consistency: Identity evidence is coherent across the full face.
B05	A live-captured headshot. Identity: Single, non-composite identity. Geometry: No inconsistent boundaries or interpolated features. Texture: Raw, unmodified texture reflecting natural variance. Consistency: All regions of the face agree on one identity.
B06	A neutral-expression face image. Identity: Distinct individual without ambiguity. Geometry: Humanly plausible and undistorted structure. Texture: Consistent skin texture with real-world variation. Consistency: No evidence of hybrid-level manipulation.
B07	A clear portrait photograph. Identity: An identity exclusively associated with one person. Geometry: Anatomically intact and proportionate. Texture: Authentic high-frequency texture remains visible. Consistency: The entire sample reflects one biological source.
B08	A high-quality biometric portrait. Identity: Not blended or shared with other identities. Geometry: Undistorted spatial alignment of facial landmarks. Texture: Detailed but organically irregular surface patterns. Consistency: Identity cues do not contradict each other.
B09	A standard ID headshot. Identity: One independent human identity. Geometry: Natural boundary definition without mismatches. Texture: Realistic texture gradients across the face. Consistency: Identity evidence is stable across all regions.
M01	A portrait of a human face. Identity: A fused identity combining two or more individuals. Geometry: Intermediate or blended alignment across facial landmarks. Texture: Softened skin lacking natural detail due to averaging. Consistency: Biometric cues indicate mixed identity origins.
M02	A synthetic identity portrait. Identity: Not attributable to one real subject. Geometry: Interpolated structure between multiple source shapes. Texture: Averaged texture statistics masking individuality. Consistency: Identity signals appear inconsistent.
M03	A digitally blended face. Identity: A composite identity formed from distinct individuals. Geometry: Ambiguous feature shapes between donors. Texture: Smoothing and blending create artificial appearance. Consistency: The face does not represent one coherent identity.
M04	A morphed biometric attack sample. Identity: Combined identity characteristics. Geometry: Feature placement lies between two human structures. Texture: Surface shows ghosting or smoothing artifacts. Consistency: Identity is fragmented and non-uniform.
M05	A hybrid identity composition. Identity: Represents a statistical mixture of identities. Geometry: Ambiguous alignment of critical feature distances. Texture: Reduced texture complexity due to fusion. Consistency: No single person can be inferred.
M06	A blended impersonation image. Identity: Derived from multiple human contributors. Geometry: Partially aligned but contradictory feature structure. Texture: Loss of fine detail typical of averaging operations. Consistency: Face signals conflict across regions.
M07	A composite synthesized identity. Identity: Neither one person nor another, but a merge. Geometry: Geometric transitions between identity sets. Texture: Homogeneous surface appearance without individuality. Consistency: Biometric meaning varies across locations.
M08	A morph-based spoof image. Identity: Encodes multiple identity sources. Geometry: Unstable spatial cues inconsistent with one subject. Texture: Blended or flattened fine structure. Consistency: Identity mapping contradicts single-person hypothesis.
M09	A hybrid synthetic face. Identity: Representing more than one biological identity. Geometry: In-between geometric shape characteristics. Texture: Low-frequency smoothing dominates texture detail. Consistency: Identity interpretation varies across patches.

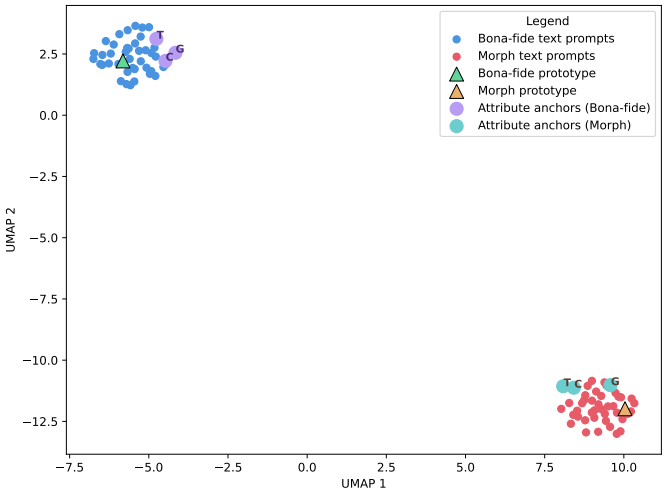
Table 2: Attribute ablation. One of four text attributes is removed during text-space learning and aligned with the image space; MAD performance is evaluated on MAD22 and MorDIFF. Best and second-best results are in **bold** and underlined, respectively.

Method	Test data	EER(%)	APCER (%) @ BPCER (%)			BPCER (%) @ APCER (%)		
			1.00	10.00	20.00	1.00	10.00	20.00
w/o I	FaceMorph	6.88	33.53	5.51	1.30	21.08	4.90	2.94
	MIPGAN I	4.36	18.10	1.70	0.70	14.22	1.96	0.98
	MIPGAN II	3.47	12.51	0.80	0.30	**8.33**	1.47	**0.00**
	OpenCV	9.28	33.94	8.64	3.96	50.98	8.33	2.94
	WebMorph	16.09	77.60	28.60	11.20	49.02	21.08	14.22
	MorDIFF	3.52	12.60	1.30	0.40	14.22	1.47	**0.00**
	Overall	7.27	31.38	7.76	2.98	26.31	6.54	3.51
w/o G	FaceMorph	5.79	28.93	3.20	0.60	16.18	3.92	1.47
	MIPGAN I	3.91	17.80	**1.20**	**0.60**	12.75	1.96	**0.49**
	MIPGAN II	3.91	11.61	0.80	**0.20**	**8.33**	1.47	**0.00**
	OpenCV	8.78	33.03	**7.62**	3.56	45.10	**5.88**	2.94
	WebMorph	15.10	76.40	26.20	8.80	48.04	17.65	13.24
	MorDIFF	3.86	12.00	1.10	0.30	12.75	1.47	**0.00**
	Overall	6.89	29.96	**6.69**	2.34	23.86	**5.39**	3.02
w/o T	FaceMorph	6.93	31.03	4.90	0.90	20.10	4.90	2.45
	MIPGAN I	4.90	17.80	1.70	0.70	13.73	1.96	0.98
	MIPGAN II	3.86	11.51	**0.70**	**0.20**	**8.33**	1.47	**0.00**
	OpenCV	**8.28**	31.61	7.83	**3.25**	50.98	**5.88**	2.94
	WebMorph	16.19	74.80	28.20	9.00	54.41	20.10	14.71
	MorDIFF	3.42	10.30	1.10	0.30	10.78	1.47	**0.00**
	Overall	7.26	29.51	7.41	2.39	26.39	5.96	3.51
w/o C	FaceMorph	5.84	30.13	3.50	0.60	16.67	4.90	1.96
	MIPGAN I	4.41	18.50	1.40	0.70	**12.25**	2.45	0.98
	MIPGAN II	3.81	12.61	0.80	0.30	8.82	1.47	**0.00**
	OpenCV	8.83	32.93	7.72	3.35	42.65	**5.88**	3.43
	WebMorph	15.64	76.80	**26.00**	9.60	53.92	20.10	13.73
	MorDIFF	3.42	11.40	1.10	0.30	11.27	1.47	**0.00**
	Overall	6.99	30.40	6.75	2.48	24.26	6.05	3.35
XSA-MAD (Ours)	FaceMorph	**4.41**	**15.72**	**2.90**	**0.40**	**14.71**	**2.45**	**0.49**
	MIPGAN I	**3.47**	**11.20**	2.20	**0.60**	14.22	**1.47**	**0.49**
	MIPGAN II	**2.92**	**8.31**	1.50	0.40	10.78	**0.49**	**0.00**
	OpenCV	10.78	**27.54**	11.89	3.86	46.57	11.27	**2.45**
	WebMorph	**13.32**	**61.00**	26.20	**5.40**	**43.14**	**16.67**	**11.27**
	MorDIFF	**1.88**	**3.00**	**0.40**	**0.10**	**2.94**	**0.00**	**0.00**
	Overall	**6.13**	**21.13**	7.52	**1.79**	**22.06**	**5.39**	**2.45**

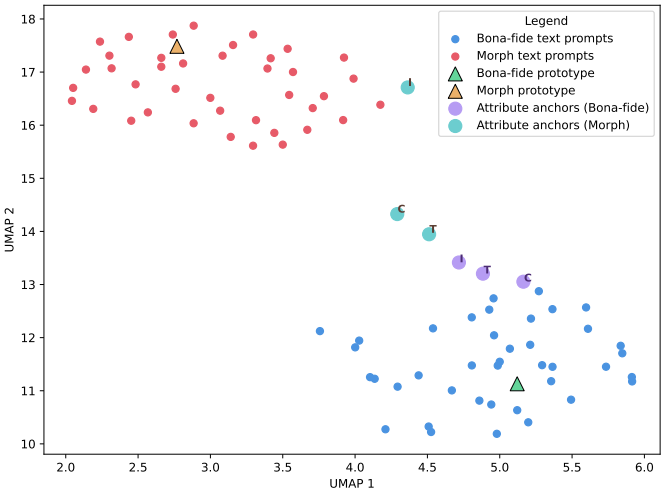


(a) default configuration

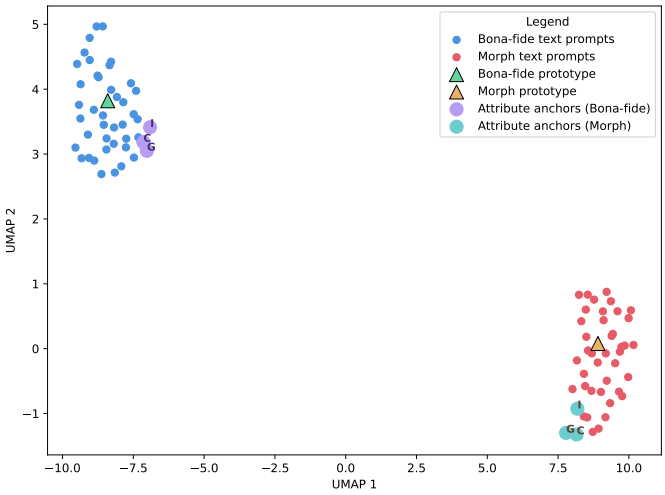
(b) Stage 1 optimized configuration



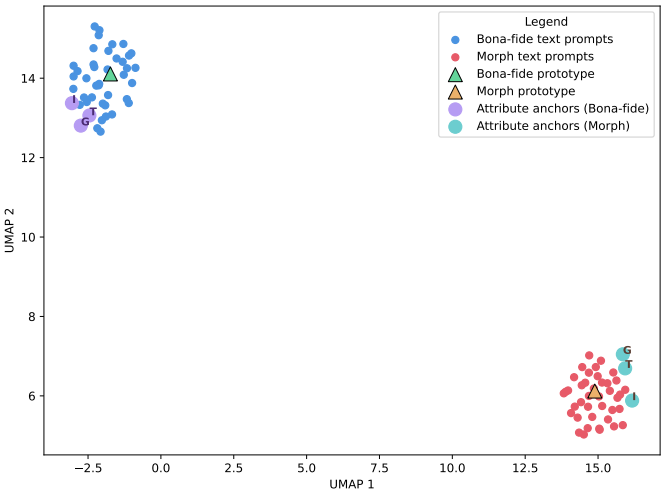
(c) w/o I configuration



(d) w/o G configuration



(e) w/o T configuration



(f) w/o C configuration

Fig. 1: UMAP visualization of the textual semantic space