Semantic Freedom Bottleneck for Domain-Generalized Multimodal Face Anti-Spoofing

Yingjie Ma[1,2,*], **Haonan Wang**[3,*], **Xun Lin**[2,4], **Ruixin Zhang**[3], **Jingyun Zhang**[3], **Jun Wang**[3], **Rizen Guo**[3], **Shouhong Ding**[3,†], **Weicheng Xie**[1,5], **Linlin Shen**[1,5,7,†], **Zitong Yu**[2,6,†]

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Institute for Artificial Intelligence, Great Bay University
[3]Tencent Youtu Lab [4]The Chinese University of Hong Kong
[5]Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University
[6]Dongguan Key Laboratory for Intelligence and Information Technology
[7]Intelligent Computing Center, Shenzhen University

Abstract

Multimodal face anti-spoofing (FAS) improves spoof detection by combining complementary RGB, infrared, and depth cues, yet its generalization to unseen domains remains fragile. We revisit multimodal domain-generalized FAS from the perspective of representation freedom. In a controlled CLIP-based post-fusion baseline, increasing visual encoder depth does not monotonically improve cross-domain performance, suggesting that higher-capacity representations may introduce additional degrees of freedom that are not necessarily aligned with real-vs-spoof semantics. To inspect this effect, we introduce the Semantic Dominance Ratio (SDR), a diagnostic statistic that measures the relative proportion of feature variation along the text-defined real-vs-spoof semantic axis versus off-axis residual directions. Motivated by this perspective, we propose Semantic Freedom Bottleneck (SFB), which reparameterizes multimodal features around CLIP text geometry. SFB decomposes each representation into a semantic component along the real-vs-spoof axis and a bounded low-rank residual component for modality-specific evidence. The residual basis is regularized to be orthogonal to the semantic axis, encouraging the residual space to retain complementary modality-specific cues without duplicating the task-semantic direction. This yields a conservative multimodal constraint that preserves discriminative real/spoof semantics while limiting unconstrained non-semantic variation. Experiments on WMCA, CASIA-CeFA, PADISI, and CASIA-SURF under fixed-modality, missing-modality, and limited-source protocols demonstrate consistently strong cross-domain performance.

1 Introduction

Face recognition is widely deployed in security-sensitive applications, yet remains vulnerable to presentation attacks (PAs) such as printed photos, replayed videos, and 3D masks. Face anti-spoofing (FAS) is therefore used as a front-end security module to reject spoofing attempts before feature matching. Although deep FAS methods perform well on standard benchmarks [53, 17, 44, 57, 41], they often degrade under unseen devices, illumination conditions, environments, and attack types, making domain generalization critical for real-world deployment. Multimodal FAS improves robustness by combining RGB, infrared (IR), and depth cues, which provide complementary

*Equal contribution.
†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

appearance, spectral, and geometric evidence [52, 10, 32, 50, 40, 45, 37, 26, 35, 36]. However, these modalities also introduce additional domain-dependent variations, including sensor style, IR response, depth noise, modality misalignment, and material-specific artifacts. Existing multimodal DG-FAS methods address these challenges through fusion, uncertainty modeling, modality rebalancing, or domain-modality alignment [27, 47]. Nevertheless, they typically do not explicitly organize representation variation around a task-defined semantic direction. As a result, high-capacity multimodal models may capture useful cross-modal evidence, but may also allocate part of their representation variation to source-dependent factors that are predictive in training domains but unreliable in unseen domains.

To inspect how such representation variation is allocated, we use a CLIP-based post-fusion baseline as a controlled diagnostic testbed. CLIP provides a shared vision-language space in which real and spoof classes can be represented by frozen text prototypes [38]. We vary the number of trainable visual encoder layers while keeping the fusion module, text prototypes, classifier, and training protocol fixed, using encoder depth as an operational intervention on visual capacity [12, 5]. We then compute the *Semantic Dominance Ratio* (SDR), a post-hoc diagnostic statistic that measures the fraction of normalized feature variation lying along the text-defined real-vs-spoof semantic axis relative to off-axis residual variation. As shown in Fig. 1, increasing encoder depth improves the ordinary CLIP baseline from 23.79% to 10.06% average HTER and from 81.55% to 95.24% average AUC, but the average SDR remains low and changes only mildly. This indicates that the performance gain does not simply come from making representation variation more dominated by the text-defined task-semantic axis. The remaining off-axis variation may contain useful modality-specific evidence, such as IR material responses or depth geometry, but may also absorb source-specific factors such as sensor style, noise statistics, or background correlations. This motivates a conservative constraint that preserves useful residual evidence while preventing unconstrained non-semantic variation from dominating the representation.

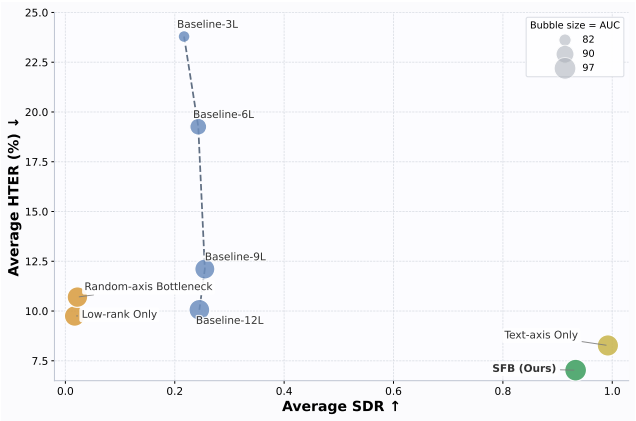


Figure 1: Layer-depth diagnosis of the ordinary CLIP post-fusion baseline under Protocol 1. Bubble size indicates AUC. Increasing the number of trainable visual encoder layers reduce HTER and improves average AUC, while SDR remains low and changes only mildly.

Motivated by this perspective, we propose *Semantic Freedom Bottleneck* (SFB) for domain-generalized multimodal face anti-spoofing. SFB uses CLIP text prototypes to construct a task-specific semantic geometry. Specifically, textual descriptions of real and spoof faces are encoded by the frozen CLIP text encoder to obtain two class prototypes, p_r and p_s, which are used as fixed prototypes for cosine-similarity classification. They also define the coordinate system of SFB: their midpoint serves as a semantic anchor c, and their normalized difference defines the real-vs-spoof semantic axis u. Given a fused multimodal feature h, SFB reparameterizes it as

$$\phi_{\mathrm{SFB}}(h) = c + \alpha(h)u + B^{\top} z(h), \tag{1}$$

where $\alpha(h)u$ is the semantic-axis component and $B^{\top} z(h)$ is a bounded low-rank residual component. The semantic component encourages the dominant variation to align with the live-vs-spoof decision, while the residual component preserves limited off-axis capacity for complementary RGB, IR, and depth evidence. Thus, SFB does not collapse the representation onto a single semantic axis; instead, it imposes a conservative inductive bias that retains bounded modality-specific evidence while restricting excessive residual freedom.

Experiments on four multimodal FAS benchmarks demonstrate that SFB achieves strong cross-domain performance under fixed-modality, missing-modality, and limited-source protocols. Our contributions are summarized as follows:

- We revisit multimodal DG-FAS from the perspective of representation variation allocation, emphasizing the need to balance modality complementarity and non-semantic residual variation.
- We introduce SDR as a post-hoc diagnostic statistic for inspecting how feature variation is distributed between a text-defined real-vs-spoof semantic axis and off-axis residual directions.
- We propose SFB, which preserves semantic-axis freedom while restricting residual freedom to a bounded low-rank subspace, providing a conservative multimodal constraint for cross-domain FAS.

2 Related Work

Cross-domain and multimodal face anti-spoofing. Face anti-spoofing (FAS) models often exploit cues entangled with illumination, sensors, environments, and attack protocols, leading to poor generalization under domain shift [53, 16, 48, 24, 25]. Existing DG-FAS methods address this issue through adversarial learning [18, 54], disentanglement [34, 19, 57], meta-learning [3, 6], data augmentation [4, 8], and feature or distribution alignment [13, 43, 22, 41]. Multimodal FAS further combines RGB, depth, and infrared cues for complementary texture, geometric, and material evidence [32, 21, 20, 52], and recent methods improve fusion with attention, adaptive supervision, cross-modal translation, flexible modality protocols, or multimodal adapters [10, 56, 31, 49–51, 28, 27]. In particular, MMDG [27] studies multimodal DG-FAS through uncertainty rectification and modality re-balancing. Unlike these methods, which mainly align domains, improve fusion, or reweight modalities, we examine the fused representation itself and analyze how its variation is allocated between a real-vs-spoof semantic direction and off-axis residual directions, connecting our motivation to shortcut learning and spurious correlations [9, 2, 23].

Vision-language guidance and semantic representation constraints. Vision-language models such as CLIP provide transferable semantic spaces and have been used in FAS through prompt-based classification, text supervision, and adapter-based adaptation [38, 7, 33, 29, 15]. Rather than using language only as auxiliary supervision or class prototypes, we use real/spoof text prototypes to define a task-specific semantic geometry for diagnosing and constraining representation variation. Our method is also related to bottleneck, low-rank, and subspace-constrained learning [42, 1, 14], as well as orthogonal subspace modeling for generalizable detection [46]. However, multimodal FAS must preserve off-axis RGB texture, IR material, and depth geometry evidence instead of treating residual variation as a purely forgery-specific branch. SFB therefore defines the real-vs-spoof decision coordinate with CLIP text geometry while retaining a bounded residual channel for complementary multimodal cues.

3 Method

3.1 Overall Framework and Feature Decomposition

We address domain-generalized multimodal face anti-spoofing (FAS), where a model trained on source domains is expected to generalize to unseen target domains. Given a face sample with available modalities $\mathcal{M}_a \subseteq \mathcal{M}$, e.g., RGB, IR, and depth, the input image of modality m is denoted as I_m. A visual encoder $E_v(\cdot)$ first extracts modality-specific representations:

$$v_m = E_v(I_m), \qquad m \in \mathcal{M}_a. \tag{2}$$

The modality features are then aggregated by a fusion operator:

$$h = \mathcal{G}_\theta(\{v_m\}_{m\in\mathcal{M}_a}), \qquad h \in \mathbb{R}^D. \tag{3}$$

We write the fusion stage as a generic operator $\mathcal{G}_\theta$ so that SFB is applied to the fused representation rather than tied to a particular fusion architecture. Under missing-modality settings, $\mathcal{G}_\theta$ is applied only to the currently available modality features.

Since the SFB module introduced in Sec. 3.2 produces a reparameterized feature $\tilde{h}$ from the fused representation h, we perform the following decomposition on $\tilde{h}$. For the i-th sample, let

$$v_i = \mathrm{Norm}(\tilde{h}_i), \tag{4}$$

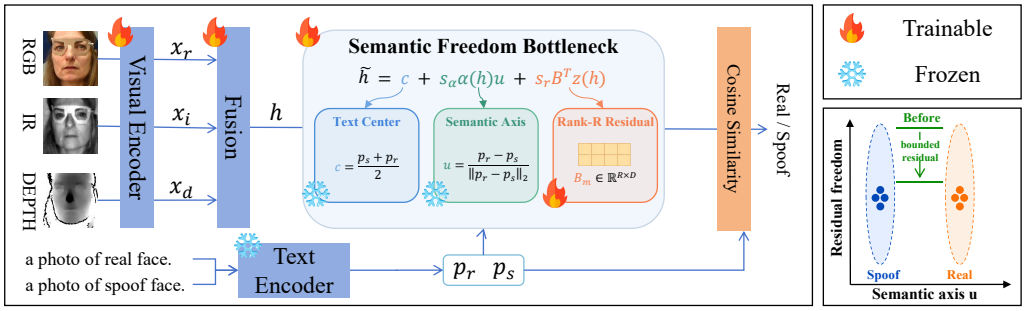


Figure 2: Overview of the proposed Semantic Freedom Bottleneck (SFB) framework. Multimodal visual features are first aggregated by a fusion operator and then reparameterized around a text-defined real-vs-spoof semantic axis, followed by prototype-based cosine classification. **Note:** Face images are from the WMCA [11] dataset.

we define its semantic-axis coordinate and off-axis residual component as

$$s_i = v_i^\top u, \qquad r_i = v_i - s_i u. \tag{5}$$

Here, s_i measures the position of the sample along the real-vs-spoof semantic axis, while r_i denotes the residual component orthogonal to that axis. Given N samples, the semantic-axis variation and off-axis residual variation are defined as

$$V_{\text{sem}} = \frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s})^2, \qquad V_{\text{res}} = \frac{1}{N}\sum_{i=1}^{N}\|r_i - \bar{r}\|_2^2. \tag{6}$$

This decomposition gives the Semantic Dominance Ratio (SDR) used in our diagnostic analysis:

$$\text{SDR} = \frac{V_{\text{sem}}}{V_{\text{sem}} + V_{\text{res}} + \epsilon}, \tag{7}$$

where ϵ is a small constant for numerical stability. SDR measures the fraction of feature variation distributed along the text-defined real-vs-spoof semantic axis, while V_{res} characterizes the scale of off-axis residual variation. SDR is used only as a post-hoc diagnostic statistic and is not used as a training objective, model-selection criterion, or threshold-tuning signal. Accordingly, a larger SDR should not be interpreted as a standalone indicator of better generalization. The trainable SFB module below imposes a related but distinct constraint: it encourages the fused representation to use the real-vs-spoof semantic axis as the primary decision coordinate while retaining a bounded off-axis residual channel.

3.2 Semantic Freedom Bottleneck

We construct the real-vs-spoof semantic geometry using a frozen CLIP text encoder. Let $\mathcal{T}_r$ and $\mathcal{T}_s$ denote prompt ensembles describing real and spoof faces, respectively. Each class prototype is obtained by averaging normalized CLIP embeddings over its prompt ensemble, whose complete templates are listed in Appendix C.4. The same prompt ensembles are fixed across all datasets, protocols, and ablations, providing a dataset-independent real-vs-spoof text contrast for SFB.

The two class text prototypes are obtained as

$$p_r = \text{Norm}\left(\frac{1}{|\mathcal{T}_r|}\sum_{\tau\in\mathcal{T}_r}\text{Norm}(E_t(\tau))\right), \tag{8}$$

$$p_s = \text{Norm}\left(\frac{1}{|\mathcal{T}_s|}\sum_{\tau\in\mathcal{T}_s}\text{Norm}(E_t(\tau))\right), \tag{9}$$

where $E_t(\cdot)$ is the frozen CLIP text encoder and $\text{Norm}(\cdot)$ denotes ℓ_2 normalization. The prototypes p_r and p_s represent real and spoof faces, respectively, and are later used as fixed class prototypes for cosine-similarity classification. They also define the semantic coordinate system of SFB:

$$c = \frac{p_s + p_r}{2}, \qquad u = \frac{p_r - p_s}{\|p_r - p_s\|_2}. \tag{10}$$

Here, c is a semantic anchor between the two class text prototypes, and u is the real-vs-spoof semantic axis pointing from spoof to real. Since both p_r and p_s are normalized, c is orthogonal to u, a property used in the margin analysis in Sec. 3.3. Note that c is not a classification prototype; the classifier uses p_s and p_r.

Given the fused multimodal feature h, SFB reparameterizes it as

$$\tilde{h} = c + s_\alpha \alpha(h) u + s_r B^\top z(h), \tag{11}$$

where $\alpha(h) \in \mathbb{R}$ is a scalar coefficient along the real-vs-spoof semantic axis, $z(h) \in \mathbb{R}^R$ is a residual coefficient vector, and $B \in \mathbb{R}^{R \times D}$ is a learnable residual basis with $R \ll D$. Therefore, $B^\top z(h)$ lies in a residual subspace whose dimension is at most R. The scalars s_α and s_r are learnable scale factors.

This reparameterization organizes the fused feature into two components. The semantic-axis component $s_\alpha \alpha(h) u$ provides the main coordinate for real-vs-spoof discrimination. The residual component $s_r B^\top z(h)$ provides a compact channel for information that may not be fully expressed by a single semantic coordinate. In multimodal FAS, such off-axis information can include complementary appearance, spectral, and geometric evidence from RGB, IR, and depth modalities. Since this channel has dimension at most R, it limits the number of off-axis directions available after fusion; the residual energy regularizer in Sec. 3.3 further controls its magnitude. Thus, SFB forms a controlled representation decomposition between the text-defined real-vs-spoof semantic axis and a low-dimensional residual subspace.

3.3 Prototype Classification and Optimization

After SFB, the constrained feature $\tilde{h}$ is classified by cosine similarity to fixed text prototypes:

$$\ell = \gamma \cdot \mathrm{Norm}(\tilde{h}) \left[\mathrm{Norm}(p_s), \mathrm{Norm}(p_r)\right]^\top, \tag{12}$$

where γ is the logit scale and $\ell \in \mathbb{R}^2$ denotes the spoof/real logits. This classifier keeps the final decision space aligned with the same text-defined semantic coordinate system used by SFB.

The prototype classifier also reveals how the semantic and residual components affect the decision margin. Since

$$p_r - p_s = \|p_r - p_s\|_2 u, \tag{13}$$

and

$$c = \frac{p_s + p_r}{2} \tag{14}$$

is orthogonal to u, the real-vs-spoof logit margin can be written as

$$\ell_r - \ell_s = \frac{\gamma \|p_r - p_s\|_2}{\|\tilde{h}\|_2} \left(s_\alpha \alpha(h) + s_r z(h)^\top B u\right). \tag{15}$$

This expression separates the direct semantic-axis contribution $s_\alpha \alpha(h)$ from the direct residual projection term $s_r z(h)^\top B u$. We constrain the residual basis to have small alignment with u, so that Bu remains small and the real-vs-spoof margin is encouraged to be mainly controlled by the semantic-axis component, while the low-dimensional residual channel can still participate in representation learning. The residual component may still affect classification indirectly through feature normalization and the preceding representation. Thus, the orthogonality constraint should be understood as limiting the residual branch from duplicating the semantic-axis margin direction, rather than removing all residual influence.

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_\perp \mathcal{L}_{orth} + \lambda_r \mathcal{L}_{res}, \tag{16}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss. The orthogonality regularizer is defined as

$$\mathcal{L}_{orth} = \frac{1}{R} \sum_{i=1}^{R} \left(\mathrm{Norm}(B_i)^\top u\right)^2, \tag{17}$$

Table 1: Cross-dataset testing results under the fixed-modal scenarios (Protocol 1) among CASIA-CeFA (C), PADISI (P), CASIA-SURF (S), and WMCA (W). Best results are marked in **bold**.

Method	CPS→W		CPW→S		CSW→P		PSW→C		Average	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
Uni-modal DG (Concat + 1×1 Conv)										
SSDG [17]	26.09	82.03	28.50	75.91	41.82	60.56	40.48	62.31	34.22	70.20
SSAN [44]	17.73	91.69	27.94	79.04	34.49	68.85	36.43	69.29	29.15	77.22
SA-FAS [41]	21.37	87.65	23.22	84.49	35.10	70.86	35.38	69.71	28.77	78.18
IADG [57]	27.02	86.50	23.04	83.11	32.06	73.83	39.24	63.68	30.34	76.78
CLIP [38]	14.55	90.47	18.17	90.02	24.13	83.15	38.33	65.71	23.80	82.34
FLIP [28]	13.19	93.79	11.73	94.93	17.39	90.63	22.14	83.95	16.11	90.83
Multi-modal FAS										
ViT [5]	20.88	84.77	44.05	57.94	33.58	71.80	42.15	56.45	35.17	67.74
AMA [50]	17.56	88.74	27.50	80.00	21.18	85.51	47.48	55.56	28.43	77.45
VP-FAS [49]	16.26	91.22	24.42	81.07	21.76	85.46	39.35	66.55	25.45	81.08
ViTAF [16]	20.58	85.82	29.16	77.80	30.75	73.03	39.75	63.44	30.06	75.02
MM-CDCN [52]	38.92	65.39	42.93	59.79	41.38	61.51	48.14	53.71	42.84	60.10
CMFL [10]	18.22	88.82	31.20	75.66	26.68	80.85	36.93	66.82	28.26	78.04
MMDG [27]	12.79	93.83	15.32	92.86	18.95	88.64	29.93	76.52	19.25	87.96
DADM [47]	11.71	94.89	**6.92**	**97.66**	19.03	88.22	16.87	91.08	13.63	92.96
SFB	**2.32**	**99.64**	8.04	96.88	**9.15**	**96.92**	**8.61**	**95.66**	**7.03**	**97.28**

where B_i is the i-th row of B. This term encourages the residual basis to be orthogonal to the real-vs-spoof semantic axis and discourages the residual channel from duplicating the main semantic direction. The residual energy penalty is

$$\mathcal{L}_{res} = \left\| B^\top z(h) \right\|_2^2, \tag{18}$$

which controls the magnitude of the off-axis residual channel. Together, these regularizers constrain the residual space so that the model remains primarily organized around the real-vs-spoof semantic axis while retaining bounded auxiliary multimodal evidence.

4 Experiments

4.1 Experimental Setup

Datasets and protocols. We evaluate the proposed SFB on four widely used multimodal face anti-spoofing benchmarks: CASIA-CeFA (C) [30], PADISI (P) [39], CASIA-SURF (S) [55], and WMCA (W) [11]. For a fair comparison with prior multimodal DG-FAS methods, we follow the dataset splits, modality settings, and evaluation protocols of MMDG [27]. We report results under three protocols: Protocol 1 evaluates complete-modality leave-one-domain-out testing (CPS→W, CPW→S, CSW→P, and PSW→C); Protocol 2 evaluates test-time missing-modality robustness, where models are trained with complete modalities but tested with missing depth, missing IR, or missing depth and IR; and Protocol 3 evaluates limited-source-domain generalization (CW→PS and PS→CW). At test time, missing modalities are skipped and fusion is applied only to the available modality features. We use HTER and AUC as the main metrics. More details on datasets and protocols are provided in Appendix A.

Implementation details. We adopt a pretrained CLIP ViT encoder as the visual backbone. All modality images are resized to 224×224 and encoded into visual representations, which are projected into a 512-dimensional feature space. RGB, IR, and depth features are aggregated by a lightweight multimodal fusion module, followed by the proposed Semantic Freedom Bottleneck. Unless otherwise specified, we use post-fusion SFB with residual rank $R = 16$. The model is trained end-to-end using AdamW with an initial learning rate of 5×10^{-6}, weight decay 1×10^{-3}, batch size 72, and 100 training epochs. The orthogonality and residual regularization weights are fixed across all protocols. Dataset statistics and protocol details are provided in Appendix A.

4.2 Cross-Dataset Testing

Fixed-modal cross-dataset testing. Table 1 reports the results under Protocol 1, where all modalities are available at both training and test time. SFB achieves the best average performance with 7.03% HTER and 97.28% AUC, outperforming the strongest competing method DADM by 6.60 percentage points in average HTER and 4.32 percentage points in average AUC. The gains are especially large on CPS→W and PSW→C, where SFB reduces HTER from 11.71% to 2.32% and from 16.87% to

Table 2: Cross-dataset testing results under the missing-modality scenarios (Protocol 2). Best results are marked in **bold**.

Method	Missing D		Missing I		Missing D & I		Average	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
Uni-modal DG (Concat + 1×1 Conv)								
SSDG [17]	38.92	65.45	37.64	66.57	39.18	65.22	38.58	65.75
SSAN [44]	36.77	69.21	41.20	61.92	33.52	73.38	37.16	68.17
SA-FAS [41]	36.30	69.07	39.80	62.69	33.08	74.29	36.39	68.68
IADG [57]	40.72	58.72	42.17	61.83	37.50	66.90	40.13	62.48
FLIP [28]	23.66	83.90	24.06	84.04	27.07	79.79	24.93	82.58
CLIP [38]	28.07	77.00	29.10	77.04	32.58	73.36	29.92	75.80
Multi-modal FAS								
ViT [5]	40.04	64.69	36.77	68.19	36.20	69.02	37.67	67.30
AMA [50]	29.25	77.70	32.30	74.06	31.48	75.82	31.01	75.86
VP-FAS [49]	29.13	78.27	29.63	77.51	30.47	76.31	29.74	77.36
ViTAF [16]	34.99	73.22	35.88	69.40	35.89	69.61	35.59	70.74
MM-CDCN [52]	44.90	55.35	43.60	58.38	44.54	55.08	44.35	56.27
CMFL [10]	31.37	74.62	30.55	75.42	31.89	74.29	31.27	74.78
MMDG [27]	24.89	82.39	23.39	83.82	25.26	81.86	24.51	82.69
DADM [47]	21.56	85.17	20.82	85.28	22.61	84.04	21.66	84.83
SFB	**18.25**	**86.59**	**9.00**	**95.95**	**17.66**	**87.39**	**14.97**	**89.98**

8.61%, respectively. These results suggest that the proposed text-axis bottleneck provides an effective cross-domain constraint beyond standard multimodal fusion and domain-generalization baselines.

Missing-modality robustness. Table 2 evaluates Protocol 2, where models are trained with complete modalities but tested with missing inputs. SFB consistently achieves the best performance across all missing-modality settings. Compared with DADM, SFB improves the average HTER from 21.66% to 14.97% and the average AUC from 84.83% to 89.98%. The improvement is particularly clear when IR is missing, where SFB obtains 9.00% HTER and 95.95% AUC. This suggests that SFB maintains a more stable fused representation when modality availability changes at test time, which is consistent with the motivation of bounding off-axis modality-dependent variation.

Limited-source-domain testing. Table 3 reports Protocol 3 results. When only two source domains are available, SFB remains competitive with state-of-the-art multimodal DG methods. On CW→PS, SFB achieves 14.10% HTER and 92.87% AUC, close to DADM. On the more challenging PS→CW setting, SFB achieves the best performance with 11.46% HTER and 95.15% AUC, substantially outperforming DADM by 8.94 percentage points in HTER and 5.64 points in AUC. These results suggest that SFB provides a useful inductive bias when source-domain supervision is limited, where conservative control of residual variation becomes especially important.

Table 3: Limited-source-domain results under Protocol 3. Best results are in **bold**.

Method	CW→PS		PS→CW	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
Uni-modal DG				
SSDG [17]	25.34	80.17	46.98	54.29
SSAN [44]	26.55	80.06	39.10	67.19
SA-FAS [41]	25.20	81.06	36.59	70.03
IADG [57]	22.82	83.85	39.70	63.46
FLIP [28]	15.92	92.38	23.85	83.46
CLIP [38]	19.36	90.57	29.98	79.22
Multi-modal FAS				
ViT [5]	42.66	57.80	42.75	60.41
AMA [50]	29.25	76.89	38.06	67.64
VP-FAS [49]	25.90	81.79	44.37	60.83
ViTAF [16]	29.64	77.36	39.93	61.31
MM-CDCN [52]	29.28	76.88	47.00	51.94
CMFL [10]	31.86	72.75	39.43	63.17
MMDG [27]	20.12	88.24	36.60	70.35
DADM [47]	**12.61**	**93.81**	20.40	89.51
SFB	14.10	92.87	**11.46**	**95.15**

4.3 Diagnostic

Analysis with Semantic Dominance Ratio

We use the SDR defined in Sec. 3.1 to inspect how feature variation is allocated between the text-defined real-vs-spoof semantic axis and off-axis residual directions. SDR is computed offline after training and is not used for optimization, checkpoint selection, or threshold tuning. We also report the off-axis freedom ratio $\mathrm{FR} = 1 - \mathrm{SDR}$ for readability.

Table 4 shows that increasing the depth of the CLIP visual encoder improves average performance, but not monotonically across all target domains. For instance, increasing the depth from 3 to 9 layers greatly improves CPS→W, while the 9-layer model performs worse than the 3-layer model on PSW→C; the best-performing depth also differs across CPW→S and CSW→P. This motivates inspecting not only classification metrics, but also how feature variation is distributed relative to the text-defined semantic axis.

Table 4: Layer-wise HTER of the ordinary CLIP post-fusion baseline under Protocol 1.

Layers	HTER(%)↓				Avg.
	CPS→W	CPW→S	CSW→P	PSW→C	
3	33.59	8.26	40.47	12.82	23.79
6	20.40	11.15	27.58	17.90	19.26
9	2.25	5.62	20.55	20.03	12.11
12	3.98	8.93	11.70	15.62	10.06

Table 5: Effect of residual freedom budget in SFB.

Residual Rank	HTER(%)↓	AUC(%)↑
0	8.27	96.34
8	8.70	96.91
16	**7.03**	**97.28**
32	10.88	94.28

Table 6: Mechanism-oriented SDR analysis under Protocol 1. Random-axis Bottleneck serves as a negative-control variant for testing whether an arbitrary fixed axis is sufficient. FR denotes the off-axis freedom ratio.

Method	Avg. SDR↑	Avg. FR↓	Avg. HTER(%)↓	Avg. AUC(%)↑
Baseline-L3	0.217	0.783	23.79	81.55
Baseline-L6	0.243	0.757	19.26	88.38
Baseline-L9	0.255	0.745	12.11	93.92
Baseline-L12	0.245	0.755	10.06	95.24
Low-rank Only	0.017	0.983	9.75	94.85
Random-axis Bottleneck	0.022	0.978	10.70	94.77
Text-axis Only	0.992	0.008	8.27	96.34
SFB	0.933	0.067	**7.03**	**97.28**

Table 6 further compares ordinary CLIP baselines with bottleneck variants. Although deeper baselines improve average HTER, their SDR remains consistently low, suggesting that increasing visual capacity does not necessarily make the representation more dominated by the real-vs-spoof semantic axis. The bottleneck variants are analyzed in Sec. 4.4.

Text-axis Only achieves very high SDR by removing the residual branch, showing that the text-defined real-vs-spoof direction provides a useful task reference. However, its performance is still worse than full SFB, showing that a higher SDR alone is not sufficient for optimal cross-domain performance. This result supports our interpretation of SDR as a diagnostic statistic rather than a standalone optimization objective. Full SFB obtains the best HTER and AUC while maintaining high, but not degenerate, semantic dominance. These results suggest that SFB is not merely a low-rank compression module or a generic axis constraint; it combines a task-relevant semantic axis with a bounded residual subspace so that dominant variation follows the real-vs-spoof semantics while complementary multimodal evidence is retained in a controlled manner.

4.4 Ablation Study

We organize the ablation study around four mechanism-level questions: whether the gain comes from generic capacity restriction, whether the bottleneck axis must be task-relevant, whether SFB simply maximizes SDR by removing all residual variation, and how much residual freedom is useful. All variants in Table 6 use the same CLIP backbone, multimodal fusion module, and training protocol, so the comparison isolates the effect of the bottleneck design. Text-axis Only sets $R = 0$ and removes the residual branch. Low-rank Only keeps a compact bottleneck without using the CLIP-defined real-vs-spoof axis as the main coordinate. Random-axis Bottleneck replaces the text-defined axis in SFB with a fixed random unit direction.

Generic capacity restriction. The Low-rank Only variant introduces a compact residual bottleneck without using the text-defined real-vs-spoof semantic axis. It improves over ordinary CLIP baselines in average HTER, suggesting that restricting excessive feature capacity can provide some regularization benefit. However, its SDR remains close to zero, indicating that most feature variation is still allocated to off-axis directions rather than the task-relevant semantic direction. This result suggests that low-rank compression can provide useful regularization, but is insufficient to align representation variation with the real-vs-spoof semantic axis. Therefore, the improvement of SFB cannot be attributed solely to dimensionality reduction or generic bottleneck regularization.

Task-relevant semantic direction. To distinguish semantic alignment from axis-based constraint itself, we introduce a Random-axis Bottleneck as a negative-control variant. This variant follows the same bottleneck principle but replaces the real-vs-spoof text axis with a task-irrelevant fixed random direction. As shown in Table 6, Random-axis Bottleneck obtains low SDR and worse average HTER than both Text-axis Only and full SFB. This comparison suggests that the gain does not come from an arbitrary one-dimensional axis constraint alone. The bottleneck direction must provide a task-relevant

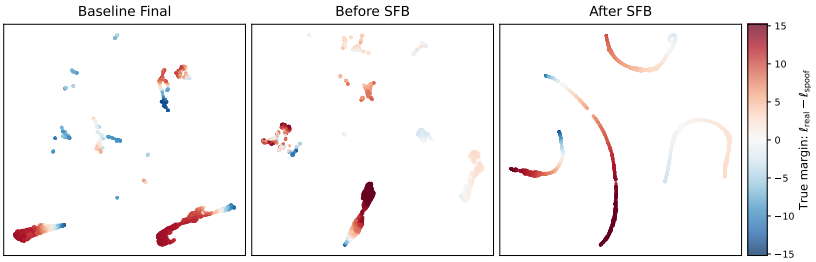


Figure 3: UMAP visualization of semantic-margin organization. Points are colored by the true real-vs-spoof margin $\ell_r - \ell_s$ with a shared color scale. Compared with the fragmented baseline and pre-SFB representations, the final SFB feature forms smoother trajectories with more continuous margin variation.

Table 7: Computational cost comparison under RGB-D-IR input. Latency is measured with batch size 1 and image size 224×224 on a single GPU.

Method	Params (M)	Trainable (M)	FLOPs (G)	Latency (ms)	Throughput	Mem. (MB)
Baseline-L12	152.775	89.347	50.614	**14.439**	**69.26**	**1104.67**
Low-rank-R16	152.793	89.364	50.614	15.268	65.50	1104.77
SFB-R16	152.794	89.366	50.614	15.117	66.15	1104.78

semantic reference for live/spoof discrimination. In this sense, the CLIP-defined real-vs-spoof axis is not merely used as an auxiliary classifier, but serves as the coordinate direction that determines where the dominant discriminative freedom should be allocated.

Non-trivial semantic dominance. The Text-axis Only variant removes the residual branch by setting $R = 0$, forcing the representation to rely only on the real-vs-spoof semantic-axis component. This variant achieves the highest SDR, showing that the text-defined semantic axis provides a useful task reference. However, Text-axis Only is still inferior to full SFB in both average HTER and AUC. This observation is important because it shows that SFB does not work by trivially maximizing SDR or by eliminating all residual variation. Instead, SDR should be interpreted as a diagnostic statistic of representation freedom allocation rather than a standalone optimization target. Robust multimodal FAS requires semantic dominance, but it should not collapse all information onto a single axis. Full SFB achieves better performance by maintaining high semantic dominance while preserving a bounded residual subspace for complementary modality-specific evidence.

Residual freedom budget. Table 5 studies the residual freedom budget controlled by rank R. When $R = 0$, the model degenerates to Text-axis Only. This setting strongly suppresses residual variation but is not optimal, suggesting that some modality-specific residual evidence remains useful for multimodal spoof detection. Increasing the rank to $R = 16$ gives the best performance, indicating that a moderate residual subspace can preserve complementary RGB, IR, and depth cues without overwhelming the task-semantic direction. However, further increasing the rank to $R = 32$ degrades performance, suggesting that simply increasing residual capacity is not beneficial and may allow more off-axis variation than needed under the cross-domain setting. These results support the central design principle of SFB: residual freedom should be retained, but explicitly bounded.

Computational cost. Table 7 compares the cost of SFB with the baseline. SFB introduces only a negligible increase in parameters and memory, while keeping FLOPs unchanged at the reported precision. The latency overhead is also small. Therefore, the performance gain of SFB is not obtained by substantially increasing model size or computation, but by imposing a semantic bottleneck on the fused representation.

Representation visualization. Fig. 3 visualizes the baseline, pre-SFB, and final SFB features with UMAP, with points colored by the true real-vs-spoof margin $\ell_r - \ell_s$. Compared with the fragmented baseline and pre-SFB features, the final SFB representation forms smoother trajectories with more continuous margin variation, supporting the semantic-residual geometry induced by SFB. Additional visualizations are provided in Appendix C.

5 Conclusion

We study cross-domain multimodal face anti-spoofing through the lens of representation freedom allocation. Using SDR as a diagnostic statistic, we show that stronger CLIP visual capacity does not necessarily make representations more dominated by the real-vs-spoof semantic direction. We then propose Semantic Freedom Bottleneck (SFB), which preserves semantic-axis freedom while restricting off-axis residual variation to a bounded low-rank subspace. Experiments on four multimodal FAS benchmarks demonstrate consistent gains under fixed-modality, missing-modality, and limited-source-domain protocols.

Limitations. SDR is intended as a diagnostic statistic rather than a selection or optimization objective. SFB relies on a fixed CLIP text-ensemble real-vs-spoof contrast and uses a fixed residual rank across protocols. Future work will explore more adaptive semantic directions and residual freedom control for broader deployment settings.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 62576076, 62276170, 62576216), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025B1515120017), Guangdong Provincial Key Laboratory (Grant No. 2023B1212060076), CCF-Tencent Rhino-Bird Open Research Fund, Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047), Intelligent Computing Center of Shenzhen University. The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University.

References

[1] Achille, A., Soatto, S.: Emergence of invariance and disentanglement in deep representations. Journal of Machine Learning Research **19**(50), 1–34 (2018)

[2] Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)

[3] Cai, R., Li, Z., Wan, R., Li, H., Hu, Y., Kot, A.C.: Learning meta pattern for face anti-spoofing. IEEE TIFS **17**, 1201–1213 (2022)

[4] Cai, R., Soh, C., Yu, Z., Li, H., Yang, W., Kot, A.C.: Towards data-centric face anti-spoofing: Improving cross-domain generalization via physics-based data synthesis. IJCV pp. 1–22 (2024)

[5] Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)

[6] Du, Z., Li, J., Zuo, L., Zhu, L., Lu, K.: Energy-based domain generalization for face anti-spoofing. In: Proceedings of the 30th ACM MM. pp. 1749–1757 (2022)

[7] Fang, H., Liu, A., Jiang, N., Lu, Q., Zhao, G., Wan, J.: Vl-fas: Domain generalization via vision-language model for face anti-spoofing. In: ICASSP 2024-2024 IEEE ICASSP. pp. 4770–4774. IEEE (2024)

[8] Ge, X., Liu, X., Yu, Z., Shi, J., Qi, C., Li, J., Kälviäinen, H.: DiffFas: face anti-spoofing via generative diffusion models. In: ECCV. pp. 144–161. Springer (2025)

[9] Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)

[10] George, A., Marcel, S.: Cross modal focal loss for rgbd face anti-spoofing. In: Proceedings of the IEEE/CVF conference on CVPR. pp. 7882–7891 (2021)

[11] George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel, S.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE TIFS **15**, 42–55 (2019)

[12] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)

[13] Hu, C., Zhang, K.Y., Yao, T., Ding, S., Ma, L.: Rethinking generalizable face anti-spoofing via hierarchical prototype-guided distribution refinement in hyperbolic space. In: Proceedings of the IEEE/CVF Conference on CVPR. pp. 1032–1041 (2024)

[14] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)

[15] Hu, X., Liu, H., Yuan, H., Fu, Z., Luo, Y., Zhang, N., Zou, H., Gan, J., Zhang, Y.: Fine-grained prompt learning for face anti-spoofing. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7619–7628 (2024)

[16] Huang, H.P., Sun, D., Liu, Y., Chu, W.S., Xiao, T., Yuan, J., Adam, H., Yang, M.H.: Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In: ECCV. pp. 37–54. Springer (2022)

[17] Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on CVPR. pp. 8484–8493 (2020)

[18] Jiang, F., Li, Q., Liu, P., Zhou, X.D., Sun, Z.: Adversarial learning domain-invariant conditional features for robust face anti-spoofing. IJCV **131**(7), 1680–1703 (2023)

[19] Jiang, F., Li, Q., Wang, W., Ren, M., Shen, W., Liu, B., Sun, Z.: Open-set single-domain generalization for robust face anti-spoofing. IJCV pp. 1–22 (2024)

[20] Kong, C., Zheng, K., Liu, Y., Wang, S., Rocha, A., Li, H.: M3-FAS: an accurate and robust multimodal mobile face anti-spoofing system. IEEE TDSC (2024)

[21] Kong, C., Zheng, K., Wang, S., Rocha, A., Li, H.: Beyond the pixel world: A novel acoustic-based face anti-spoofing system for smartphones. IEEE TIFS **17**, 3238–3253 (2022)

[22] Le, B.M., Woo, S.S.: Gradient alignment for cross-domain face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on CVPR. pp. 188–199 (2024)

[23] Li, B., Shen, Y., Wang, Y., Zhu, W., Li, D., Keutzer, K., Zhao, H.: Invariant information bottleneck for domain generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 7399–7407 (2022)

[24] Li, Z., Zhao, T., Xu, X., Zhang, Z., Li, Z., Chen, X., Zhang, Q., Bergamo, A., Jain, A.K., Xing, Y.: Optimal transport-guided source-free adaptation for face anti-spoofing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24351–24363 (2025)

[25] Lin, K.H., Tseng, Y.W., Huang, K.Y., Wu, J.C., Cheng, W.H.: Instructflip: Exploring unified vision-language model for face anti-spoofing. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 2987–2996 (2025)

[26] Lin, X., Liu, A., Yu, Z., Cai, R., Wang, S., Yu, Y., Wan, J., Lei, Z., Cao, X., Kot, A.: Reliable and balanced transfer learning for generalized multimodal face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(9), 7608–7625 (2025)

[27] Lin, X., Wang, S., Cai, R., Liu, Y., Fu, Y., Tang, W., Yu, Z., Kot, A.: Suppress and rebalance: Towards generalized multi-modal face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on CVPR. pp. 211–221 (2024)

[28] Liu, A., Liang, Y.: Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. arXiv:2304.07549 (2023)

[29] Liu, A., Ma, H., Zheng, J., Yuan, H., Yu, X., Liang, Y., Escalera, S., Wan, J., Lei, Z.: Fm-clip: Flexible modal clip for face anti-spoofing. In: Proceedings of the 32nd ACM MM. pp. 8228–8237 (2024)

[30] Liu, A., Tan, Z., Wan, J., Escalera, S., Guo, G., Li, S.Z.: Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1179–1187 (2021)

[31] Liu, A., Tan, Z., Wan, J., Liang, Y., Lei, Z., Guo, G., Li, S.Z.: Face anti-spoofing via adversarial cross-modality translation. IEEE TIFS **16**, 2759–2772 (2021)

[32] Liu, A., Tan, Z., Yu, Z., Zhao, C., Wan, J., Liang, Y., Lei, Z., Zhang, D., Li, S.Z., Guo, G.: Fm-vit: Flexible modal vision transformers for face anti-spoofing. IEEE TIFS **18**, 4775–4786 (2023)

[33] Liu, A., Xue, S., Gan, J., Wan, J., Liang, Y., Deng, J., Escalera, S., Lei, Z.: Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on CVPR. pp. 222–232 (2024)

[34] Liu, Y., Li, Z., Xu, Y., Guo, Z., Zou, Z., Wu, L.: Quality-invariant domain generalization for face anti-spoofing. IJCV pp. 1–16 (2024)

[35] Ma, Y., Lin, X., Xu, Y., Xie, W., Yu, Z.: Pa-fas: towards interpretable and generalizable multimodal face anti-spoofing via path-augmented reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 7856–7864 (2026)

[36] Ma, Y., Lin, X., Yu, Z., Wang, H., Zhang, R., Ding, S., Liu, X., Yuan, X., Xie, W., Shen, L.: Purify then guide: Rethinking domain generalization for multimodal face anti-spoofing. In: European Conference on Computer Vision. pp. 674–691. Springer (2026)

[37] Ma, Y., Yu, Z., Lin, X., Xie, W., Shen, L.: Big-MOE: bypassing isolated gating for generalized multimodal face anti-spoofing. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5 (2025)

[38] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)

[39] Rostami, M., Spinoulas, L., Hussein, M., Mathai, J., Abd-Almageed, W.: Detection and continual learning of novel face presentation attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14851–14860 (2021)

[40] Shi, Y., Gao, Y., Lai, Y., Wang, H., Feng, J., He, L., Wan, J., Chen, C., Yu, Z., Cao, X.: Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. Visual Intelligence **3**(1), 9 (2025)

[41] Sun, Y., Liu, Y., Liu, X., Li, Y., Chu, W.S.: Rethinking domain generalization for face anti-spoofing: Separability and alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24563–24574 (2023)

[42] Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)

[43] Wang, K., Zhang, G., Yue, H., Liang, Y., Huang, M., Zhang, G., Han, J., Ding, E., Wang, J.: CSDG-FAS: closed-space domain generalization for face anti-spoofing. IJCV pp. 1–14 (2024)

[44] Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., Wang, Z.: Domain generalization via shuffled style assembly for face anti-spoofing. In: Proceedings of the IEEE/CVF conference on CVPR. pp. 4123–4133 (2022)

[45] Xie, X., Cui, Y., Tan, T., Zheng, X., Yu, Z.: Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. Visual Intelligence **2**(1), 37 (2024)

[46] Yan, Z., Wang, J., Jin, P., Zhang, K.Y., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Orthogonal subspace decomposition for generalizable AI-generated image detection. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J. (eds.) Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 70268–70288. PMLR (13–19 Jul 2025), https://proceedings.mlr.press/v267/yan25b.html

[47] Yang, J., Lin, X., Yu, Z., Zhang, L., Liu, X., Li, H., Yuan, X., Cao, X.: DADM: dual alignment of domain and modality for face anti-spoofing. arXiv preprint arXiv:2503.00429 (2025)

[48] Yu, J., Kim, S., Lee, K., Kwon, T., Shin, W.Y., Kim, H.Y.: Multi-view slot attention using paraphrased texts for face anti-spoofing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21117–21128 (2025)

[49] Yu, Z., Cai, R., Cui, Y., Liu, A., Chen, C.: Visual prompt flexible-modal face anti-spoofing. TDSC (2024)

[50] Yu, Z., Cai, R., Cui, Y., Liu, X., Hu, Y., Kot, A.C.: Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. IJCV pp. 1–22 (2024)

[51] Yu, Z., Liu, A., Zhao, C., Cheng, K.H., Cheng, X., Zhao, G.: Flexible-modal face anti-spoofing: A benchmark. In: Proceedings of the IEEE/CVF Conference on CVPR. pp. 6346–6351 (2023)

[52] Yu, Z., Qin, Y., Li, X., Wang, Z., Zhao, C., Lei, Z., Zhao, G.: Multi-modal face anti-spoofing based on central difference networks. In: Proceedings of the IEEE/CVF Conference on CVPR Workshops. pp. 650–651 (2020)

[53] Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., Zhao, G.: Deep learning for face anti-spoofing: A survey. IEEE TPAMI **45**(5), 5609–5631 (2022)

[54] Yue, H., Wang, K., Zhang, G., Feng, H., Han, J., Ding, E., Wang, J.: Cyclically disentangled feature translation for face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3358–3366 (2023)

[55] Zhang, S., Liu, A., Wan, J., Liang, Y., Guo, G., Escalera, S., Escalante, H.J., Li, S.Z.: Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science **2**(2), 182–193 (2020)

[56] Zheng, G., Liu, Y., Dai, W., Li, C., Zou, J., Xiong, H.: Towards unified representation of invariant-specific features in missing modality face anti-spoofing. In: ECCV. pp. 93–110 (2025)

[57] Zhou, Q., Zhang, K.Y., Yao, T., Lu, X., Yi, R., Ding, S., Ma, L.: Instance-aware domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on CVPR. pp. 20453–20463 (2023)

A Datasets, Protocols, and Metrics

A.1 Datasets

Table A8: Summary of the public datasets used in our experiments.

Dataset	Year	Training (Real/Fake)	Testing (Real/Fake)	Scenarios	Attack Types
WMCA	2019	1200 / 3600	1200 / 5539	6 types	Print, replay, occlusion, masks, etc.
SURF	2019	900 / 5397	1800 / 10764	1 type	Print (paper, cloth)
PADISI	2021	698 / 598	363 / 304	1 type	Print, replay, masks, occlusion
CeFA	2020	1301 / 1701	3600 / 16617	3 types	Print, mask, replay

We adopt four public datasets (CeFA, PADISI, SURF, and WMCA) as our training and evaluation corpora. Collected with real sensors under diverse and realistic conditions, these datasets contain bonafide samples and presentation attacks carried out with real-world materials (e.g., printed photos, replay devices, masks). RGB, IR, and DEPTH modalities are captured in settings that approximate deployment. These datasets are widely used in multimodal presentation attack detection and are regarded as representative of real-world conditions. Table A8 summarizes their main characteristics.

A.2 Evaluation Protocols

To thoroughly evaluate the model, we follow the three protocols proposed in MMDG [27]. They assess cross-domain performance under different deployment conditions, including complete modalities, missing modalities at test time, and limited source domains.

Protocol 1: Complete modalities. Multimodal FAS datasets are collected across diverse scenarios and devices, which leads to clear distribution shifts. This protocol evaluates the model through four multimodal leave-one-out (LOO) subprotocols built from CeFA (**C**), PADISI (**P**), SURF (**S**), and WMCA (**W**). For example, $CPS \rightarrow W$ means training on **C**, **P**, and **S** and testing on **W**. These settings measure performance on unseen domains.

Protocol 2: Missing modalities at test time. In practical deployments, sensor failures or network issues may cause some modalities to be unavailable. This protocol extends Protocol 1 by covering cases where one or more modalities are missing during testing, including removal of the depth modality, the infrared modality, or both. It evaluates the model's robustness to incomplete inputs.

Protocol 3: Limited source domains. Because collecting multimodal FAS data is costly, this protocol restricts the number and size of training datasets to simulate scarce data. It includes two subprotocols that limit the number of source domains: $CW \rightarrow PS$ and $PS \rightarrow CW$. The goal is to test generalization when training data are limited and to verify reliability in realistic application scenarios.

B Additional Clarifications and Derivations

This section provides additional clarification for the geometric construction, the SDR diagnostic, and the margin behavior of SFB. These discussions are intended to make explicit the assumptions and consequences of the formulation in Sec. 3. They do not introduce additional training objectives or model-selection criteria.

B.1 Text Geometry and Orthogonality of the Semantic Anchor

SFB uses two normalized CLIP text prototypes, p_r and p_s, to construct the real-vs-spoof semantic geometry:

$$c = \frac{p_s + p_r}{2}, \qquad u = \frac{p_r - p_s}{\|p_r - p_s\|_2}. \tag{19}$$

Here, p_r and p_s are the fixed class prototypes used for cosine-similarity classification, while c is only a semantic anchor used in the SFB reparameterization.

Because both p_r and p_s are ℓ_2-normalized, we have

$$c^\top u = \left(\frac{p_s + p_r}{2}\right)^\top \frac{p_r - p_s}{\|p_r - p_s\|_2} = \frac{\|p_r\|_2^2 - \|p_s\|_2^2}{2\|p_r - p_s\|_2} = 0. \tag{20}$$

Thus, the semantic anchor c is orthogonal to the real-vs-spoof axis u. This property is useful because it separates the center of the two text prototypes from the discriminative direction between them. In other words, shifting features around c does not itself create a real-vs-spoof margin; the margin is governed by the coordinate along u and by any residual component that projects onto u.

B.2 SDR as a Diagnostic of Variation Allocation

In Sec. 3.1, we decompose each normalized feature $v_i = \mathrm{Norm}(h_i)$ into its semantic-axis coordinate and off-axis residual component:

$$s_i = v_i^\top u, \qquad r_i = v_i - s_i u. \tag{21}$$

Since u is normalized, the residual component is orthogonal to the semantic axis:

$$r_i^\top u = (v_i - s_i u)^\top u = v_i^\top u - s_i \|u\|_2^2 = s_i - s_i = 0. \tag{22}$$

Therefore, r_i captures the part of the normalized feature that is outside the real-vs-spoof semantic direction.

Given N samples, the semantic-axis variation and residual variation are defined as

$$V_{\mathrm{sem}} = \frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s})^2, \qquad V_{\mathrm{res}} = \frac{1}{N}\sum_{i=1}^{N}\|r_i - \bar{r}\|_2^2. \tag{23}$$

The Semantic Dominance Ratio is

$$\mathrm{SDR} = \frac{V_{\mathrm{sem}}}{V_{\mathrm{sem}} + V_{\mathrm{res}} + \epsilon}. \tag{24}$$

SDR measures how much of the normalized feature variation lies along the text-defined real-vs-spoof semantic axis. It is not an accuracy metric: it does not use prediction thresholds, does not count correct or incorrect decisions, and does not replace HTER or AUC. It is only a representation-level diagnostic used to inspect how feature variation is allocated with respect to the chosen text-defined task axis.

This distinction is important. A high SDR indicates that feature variation is concentrated near the semantic axis, but it does not imply that the model is optimal. Multimodal FAS also benefits from off-axis evidence, such as RGB texture details, IR material response, and depth geometry. Conversely, a low SDR does not necessarily imply poor classification accuracy, because some off-axis cues may still be predictive in a particular domain. For this reason, SDR should be interpreted jointly with HTER/AUC and mechanism-oriented ablations, rather than as a standalone objective.

B.3 Why SDR Is Not Used as a Training Objective

Although SDR is useful for diagnosing feature geometry, directly optimizing it would be undesirable for two reasons.

First, maximizing SDR alone may collapse useful multimodal evidence. The numerator V_{sem} only measures variation along the one-dimensional real-vs-spoof axis, while the denominator includes all off-axis residual variation. A model can increase SDR by suppressing residual variation, but this

may remove complementary cues that are important for multimodal FAS. This is why SFB retains a residual channel instead of forcing all information onto the semantic axis.

Second, SDR is computed over a set of samples and describes dataset-level variation. It is therefore more suitable as an offline diagnostic than as an instance-wise training loss. In contrast, SFB imposes an architectural and regularization-based constraint at the feature level: it explicitly parameterizes the representation with a semantic-axis component and a bounded residual component, and then optimizes the model using cross-entropy together with the orthogonality and residual-energy regularizers.

Thus, SDR and SFB are related but not identical. SDR diagnoses how representation variation is distributed after training, while SFB constrains the representation during training through the parameterization

$$\tilde{h} = c + s_\alpha \alpha(h) u + s_r B^\top z(h). \tag{25}$$

B.4 Bounded Residual Freedom

The residual component in SFB is

$$e(h) = B^\top z(h), \qquad B \in \mathbb{R}^{R \times D}, \quad z(h) \in \mathbb{R}^R. \tag{26}$$

Since $e(h)$ lies in the span of the rows of B, its dimension is at most R. Therefore, the rank parameter R controls the number of off-axis directions available to the residual channel. This gives a dimensional bound on residual freedom.

In addition, SFB uses the residual energy penalty

$$\mathcal{L}_{res} = \|B^\top z(h)\|_2^2, \tag{27}$$

which controls the magnitude of the residual component. Therefore, the residual branch is constrained in two complementary ways: R limits the number of residual directions, and $\mathcal{L}_{res}$ limits the energy assigned to those directions.

This design is different from removing residual information. The residual channel is retained because a single scalar coordinate $\alpha(h)$ along the real-vs-spoof axis may not fully capture all useful multimodal evidence. Instead, SFB keeps the residual channel compact and bounded, so that off-axis information can assist representation learning without becoming an unconstrained source of domain-specific variation.

B.5 Effect of Orthogonality on the Prototype Margin

The prototype classifier computes logits by cosine similarity between the constrained feature $\tilde{h}$ and the fixed text prototypes p_s, p_r:

$$\ell = \gamma \cdot \mathrm{Norm}(\tilde{h}) \left[\mathrm{Norm}(p_s), \mathrm{Norm}(p_r)\right]^\top. \tag{28}$$

Because p_r and p_s are normalized and

$$p_r - p_s = \|p_r - p_s\|_2 u, \tag{29}$$

the real-vs-spoof logit margin is

$$\ell_r - \ell_s = \gamma\, \mathrm{Norm}(\tilde{h})^\top (p_r - p_s) = \frac{\gamma \|p_r - p_s\|_2}{\|\tilde{h}\|_2} \tilde{h}^\top u. \tag{30}$$

Substituting the SFB parameterization,

$$\tilde{h} = c + s_\alpha \alpha(h) u + s_r B^\top z(h), \tag{31}$$

and using $c^\top u = 0$, we obtain

$$\ell_r - \ell_s = \frac{\gamma \|p_r - p_s\|_2}{\|\tilde{h}\|_2} \left(s_\alpha \alpha(h) + s_r z(h)^\top B u \right). \tag{32}$$

This expression shows that the residual channel affects the real-vs-spoof margin directly only through the projection term $z(h)^\top B u$. The orthogonality regularizer

$$\mathcal{L}_{orth} = \frac{1}{R} \sum_{i=1}^{R} \left(\mathrm{Norm}(B_i)^\top u \right)^2 \tag{33}$$

discourages the residual basis from aligning with the semantic axis, thereby reducing this direct residual contribution to the prototype margin.

The residual component can still affect classification indirectly through the normalized feature magnitude $\|\tilde{h}\|_2$ and through the learned upstream representation. This is intended: the residual channel remains part of the representation. The purpose of $\mathcal{L}_{orth}$ is not to make the residual useless, but to prevent it from duplicating the real-vs-spoof semantic direction as an unconstrained additive margin component.

B.6 Role of the Residual Energy Penalty

The orthogonality regularizer controls the direction of the residual basis relative to the semantic axis, but it does not control the magnitude of the residual component. A residual vector can be nearly orthogonal to u while still having large norm. Such large off-axis variation may dominate the normalized feature geometry and introduce unstable domain- or modality-specific patterns.

The residual energy penalty addresses this issue:

$$\mathcal{L}_{res} = \|B^\top z(h)\|_2^2. \tag{34}$$

Together with the rank-bounded residual parameterization, this penalty discourages excessive off-axis variation. Therefore, $\mathcal{L}_{orth}$ and $\mathcal{L}_{res}$ play different roles: $\mathcal{L}_{orth}$ discourages the residual basis from duplicating the semantic-axis direction, while $\mathcal{L}_{res}$ controls the overall residual magnitude.

B.7 Interpretation of Special Cases

SFB includes several meaningful special cases.

Text-axis only. When the residual rank is set to $R = 0$, the residual component is removed and the representation becomes

$$\tilde{h} = c + s_\alpha \alpha(h) u. \tag{35}$$

This forces all instance-dependent variation to lie along the real-vs-spoof semantic axis. It is useful for testing whether the text-defined axis alone is sufficient. However, this setting may discard modality-specific evidence that cannot be represented by a single scalar coordinate.

Low-rank residual without semantic axis. A generic low-rank bottleneck can restrict representation capacity, but it does not specify which direction should carry the real-vs-spoof decision. Such a variant tests whether the benefit comes merely from dimensionality reduction. In contrast, SFB explicitly anchors the main decision coordinate to the text-defined real-vs-spoof axis.

Random-axis bottleneck. Replacing u with a fixed random direction tests whether an arbitrary axis constraint is sufficient. If the random-axis variant performs worse than SFB, it supports the role of the text-defined real-vs-spoof axis rather than a generic one-dimensional bottleneck.

These special cases help separate the effects of semantic-axis anchoring, residual capacity, and generic dimensional restriction.

B.8 Implementation Notes for SDR

Unless otherwise specified, SDR is computed on the feature used by the prototype classifier. For SFB models, this corresponds to the constrained feature after SFB. For ordinary CLIP post-fusion baselines, this corresponds to the fused feature before prototype classification.

All features are ℓ_2-normalized before computing SDR. This matches the cosine-similarity classifier and removes arbitrary feature-norm scaling from the diagnostic. The residual variation can be computed without forming a full covariance matrix:

$$V_{\text{res}} = \frac{1}{N} \sum_{i=1}^{N} \|r_i - \bar{r}\|_2^2. \tag{36}$$

This is equal to the trace of the residual covariance and avoids eigendecomposition or singular-value decomposition. SDR is computed offline after training and is not used to update parameters, choose checkpoints, or tune decision thresholds.

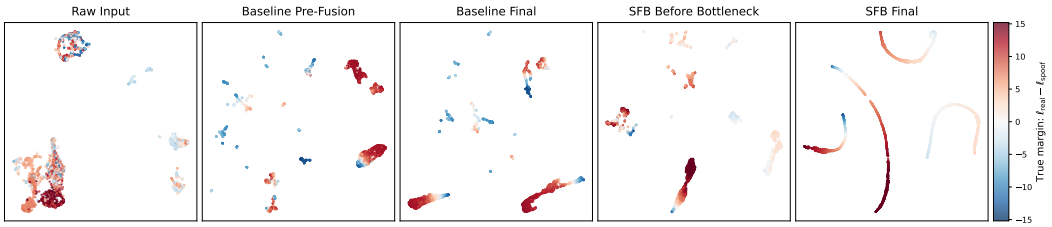


Fig. A4: UMAP visualization of the full representation pipeline, colored by the true real-vs-spoof margin $\ell_r - \ell_s$. From left to right: Raw Input, Baseline Pre-Fusion, Baseline Final, SFB Before Bottleneck, and SFB Final. A shared color scale is used across all panels. The visualization shows that while the raw and baseline representations exhibit fragmented local structures, the final SFB representation is reorganized into smoother low-dimensional trajectories with more continuous margin variation.

C Additional Representation Visualizations

To complement the mechanism analysis in the main paper, we provide additional UMAP visualizations of the representation pipeline at different stages. These visualizations are intended as qualitative evidence for how SFB reorganizes multimodal features. They do not, by themselves, prove the elimination of domain shortcuts; rather, they help reveal how the feature geometry changes from the raw input to the final representation.

C.1 Visualization Setup

We visualize five representation stages:

1. **Raw Input**: the pooled-and-flattened multimodal input;
2. **Baseline Pre-Fusion**: the concatenated modality-specific encoder outputs before fusion;
3. **Baseline Final**: the final baseline feature used for classification;
4. **SFB Before Bottleneck**: the fused representation before entering the semantic-freedom bottleneck;
5. **SFB Final**: the final representation after the bottleneck.

For each stage, we subsample the test features across the four leave-one-domain-out evaluations and project them to two dimensions with UMAP for visualization. In Fig. A4, points are colored by the true real-vs-spoof margin, i.e.,

$$\ell_r - \ell_s, \tag{37}$$

where ℓ_r and ℓ_s denote the real and spoof logits, respectively. A shared color scale is used across all panels. In Fig. A5, points are colored by the target dataset identity.

We emphasize that UMAP is used only for qualitative inspection. Because each stage is projected separately, the absolute coordinates across different panels are not directly comparable. Therefore, the interpretation should focus on within-panel structure, such as fragmentation, smoothness, and how the margin changes along the induced manifold.

C.2 Qualitative Observations

Several qualitative patterns can be observed.

Raw input already contains strong source-dependent structure. As shown in Fig. A5, the raw pooled multimodal input is already strongly grouped by dataset. This is expected, since the four datasets differ in sensor properties, collection environments, subject populations, and attack conditions. Therefore, the model is not starting from a source-invariant input space.

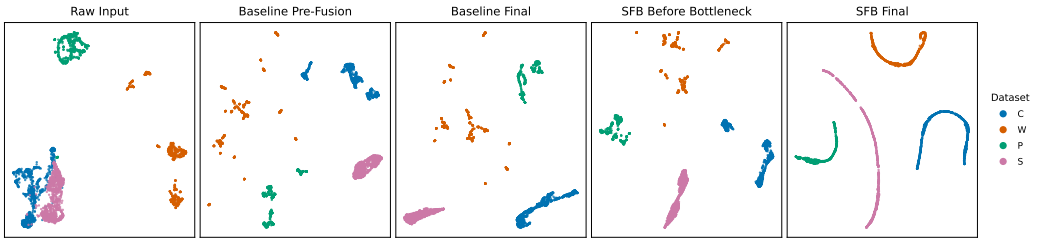


Fig. A5: UMAP visualization of the full representation pipeline, colored by target dataset identity. From left to right: Raw Input, Baseline Pre-Fusion, Baseline Final, SFB Before Bottleneck, and SFB Final. Different datasets remain visually distinguishable throughout the pipeline, including after SFB, indicating that SFB does not simply erase all source-dependent variation. Instead, it preserves source-related residual structure while reorganizing the overall representation geometry.

The baseline remains relatively fragmented after fusion. In the baseline pipeline, the pre-fusion representation already exhibits several disconnected clusters, and the baseline final representation remains largely fragmented. In Fig. A4, the true margin in the baseline final feature tends to vary locally within different feature islands rather than being organized into a globally smooth structure. This suggests that the ordinary fusion pipeline does not strongly regularize the geometry of the learned representation.

Before the bottleneck, SFB still inherits source-structured fused features. The "SFB Before Bottleneck" panel shows that the fused representation entering SFB is still structured and dataset-dependent. This is consistent with the intention of SFB: it is not designed to remove all off-axis information before the bottleneck, but rather to regularize how such information is expressed in the final representation.

After SFB, the representation is reorganized into smoother trajectories. The most striking change appears in the final SFB representation. In Fig. A4, the samples are no longer arranged as fragmented islands, but instead lie on several smooth low-dimensional trajectories. Along these trajectories, the real-vs-spoof margin varies more continuously. This supports the interpretation that SFB reshapes the fused representation into a more structured semantic-residual manifold, rather than merely improving local class separation.

SFB does not simply collapse dataset structure. Fig. A5 shows that different datasets remain distinguishable even after SFB. Therefore, the role of SFB is not to enforce complete source collapse. Instead, SFB appears to preserve source-dependent residual variation while constraining it into a more regularized and lower-freedom geometric organization.

C.3 Interpretation

Taken together, these visualizations support the following interpretation. The raw multimodal input and the ordinary baseline pipeline both preserve substantial source-dependent variation, but the resulting geometry remains fragmented. In contrast, SFB does not erase source-related structure; rather, it reorganizes the fused feature space so that the decision-related real-vs-spoof margin is arranged more continuously along smoother trajectories. This is consistent with the design goal of SFB: to constrain multimodal residual freedom while retaining useful off-axis evidence.

We stress again that these visualizations are qualitative and should be interpreted together with the quantitative performance results and the mechanism-oriented analyses in the main paper.

C.4 Prompt Ensemble for Text Prototype Construction

SFB does not construct the real-vs-spoof semantic axis from a single hand-written prompt. Instead, we use a class-wise prompt ensemble to obtain more stable CLIP text prototypes. Specifically, the real and spoof text prototypes are computed from two prompt sets, T_r and T_s, each containing 24 templates. Each template is encoded by the frozen CLIP text encoder, normalized individually, and

Table A9: Prompt ensemble used to construct the real and spoof CLIP text prototypes. The same prompt ensemble is fixed across all datasets, protocols, and ablations.

Real templates	**Spoof templates**
This is a photo of a real face	This is a photo of a spoof face
This is a photo of an authentic face	This is a photo of an attack face
This is a photo of a vital face	This is a photo of a fake face
This is a photo of a biometric face	This is a photo of a non-real face
This is a photo of a genuine face	This is a photo of a counterfeit face
This is a photo of an alive face	This is a photo of a printed spoof face
This is a photo of a real-time face	This is a photo of an adversarial patch face
This is a photo of a vivacious face	This is a photo of a wearable adversarial sticker face
This is a photo of a veritable face	This is a photo of a printed adversarial image face
This is a photo of a valid face	This is a photo of a physical impersonation attack face
This is a photo of an actual face	This is a photo of an adversarial perturbation face
This is a photo of an existent face	This is a photo of an adversarial makeup face
This is a photo of a true face	This is a photo of a facial attribute modification face
This is a photo of a legitimate face	This is a photo of an adversarial face image
This is a photo of a natural face	This is a photo of a physical-world printed attack face
This is a photo of an unfeigned face	This is a photo of a synthetic image face
This is a photo of an unimpeachable face	This is a photo of an adversarial face mask
This is a photo of a credible face	This is a photo of a semantic adversarial attack face
This is an example of a real face	This is an example of a spoof face
This is a bonafide face	This is an example of an attack face
This is a real face	This is not a real face
This is how a real face looks like	This is how a spoof face looks like
a photo of a real face	a photo of a spoof face
This is not a spoof face	a printout shown to be a spoof face

then averaged within the corresponding class:

$$p_r = \mathrm{Norm}\left(\frac{1}{|T_r|}\sum_{\tau\in T_r}\mathrm{Norm}(E_t(\tau))\right), \qquad p_s = \mathrm{Norm}\left(\frac{1}{|T_s|}\sum_{\tau\in T_s}\mathrm{Norm}(E_t(\tau))\right). \tag{38}$$

The resulting real-vs-spoof direction

$$u = \frac{p_r - p_s}{\|p_r - p_s\|_2} \tag{39}$$

is therefore an ensemble-defined CLIP real-vs-spoof contrast axis, rather than an axis induced by a single textual phrase.

The prompt ensemble is fixed once and used for all datasets, protocols, and ablations. We do not tune prompts for a particular source or target domain. The real-face templates cover generic real, authentic, genuine, live, bonafide, and non-spoof descriptions. The spoof-face templates cover generic spoof, fake, attack, non-real, printed, physical-world, mask-related, synthetic, and adversarial presentation descriptions. This construction is intended to reduce dependence on any individual wording while preserving the same real-vs-spoof semantic contrast.

Table A9 lists the complete prompt ensemble used in our experiments.

We emphasize that this prompt ensemble is used to instantiate a fixed semantic contrast, not as a tunable source of supervision. The goal is not to claim complete prompt invariance, but to avoid making the semantic coordinate system depend on a single manually selected sentence. Prompt design remains a limitation, and more systematic prompt marginalization or prompt learning may further reduce this dependency.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions and intended scope, and the claims are aligned with the presented results and stated assumptions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a discussion of limitations, including assumptions, scope of evaluation, and factors that may affect generalization or scalability.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the datasets, data splits, baselines, hyperparameters, and other experimental details needed to understand and approximately reproduce the main empirical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper describes the experimental procedure, but the code and/or data are not released openly at submission time. Reproducibility is supported through the methodological and implementation details provided in the paper and supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the dataset names, data splits, baselines, hyperparameters, and other training/test details necessary to understand the reported results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper does not report error bars, standard deviations, confidence intervals, statistical significance tests, or results over multiple random seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not provide sufficient information about compute resources such as GPU/CPU type, memory, execution time, or total compute required to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics; it does not involve human subjects, does not release high-risk assets, and uses public datasets appropriately.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential positive impacts of the work as well as possible negative impacts, misuse risks, and limitations relevant to responsible deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models that pose a high risk of misuse requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses public datasets and credits the original creators or sources; the corresponding licenses or terms of use should be explicitly stated and respected in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new assets such as datasets, code packages, or model checkpoints requiring separate documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subjects research, so IRB approval or equivalent review is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as an important, original, or non-standard component of the core research methodology; any use for writing, editing, or formatting does not require declaration under this checklist item.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.