Purify then Guide: Rethinking Domain Generalization for Multimodal Face Anti-Spoofing

Yingjie Ma[1,2†], Xun Lin[2,8†], Zitong Yu[2,3⋆], Haonan Wang[5], Ruixin Zhang[5⋆], Shouhong Ding[5], Xin Liu[6], Xiaochen Yuan[7], Weicheng Xie[1,4], and Linlin Shen[1,4⋆]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] School of Computing and Information Technology, Great Bay University
[3] Dongguan Key Laboratory for Intelligence and Information Technology
[4] Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University
[5] Tencent Youtu Lab
[6] School of Psychology, Shanghai Jiao Tong University
[7] Faculty of Applied Sciences, Macao Polytechnic University
[8] The Chinese University of Hong Kong
[9] https://github.com/murInJ/MMDA

Abstract. Face anti-spoofing (FAS) is critical for secure deployment of face recognition in high-stakes applications. However, existing multimodal FAS methods still generalize poorly to unseen domains. We argue that a key reason is that they directly align noisy multimodal features in which spoof-relevant cues are entangled with modality- and domain-specific variations, so both useful and nuisance patterns are forced to match across domains. Motivated by this, we design **M**ultimodal **D**enoising and **A**lignment (**MMDA**), a CLIP-based framework that (i) purifies multimodal features, (ii) softly guides them into a semantic space, and (iii) preserves the pre-trained geometry during task adaptation. To decouple spoof cues from these variations before alignment, the **M**odality-**D**omain Joint **D**ifferential **A**ttention (**MD2A**) module contrasts same-domain cross-sample features to down-weight patterns shared within each domain/modality, which helps reduce domain- and modality-related biases while better preserving discriminative spoof information. To avoid destructive, over-rigid alignment, the **R**epresentation **S**pace **S**oft (**RS2**) strategy aligns visual features to real/spoof subspaces spanned by multiple prompts in the CLIP space, providing a soft semantic pull instead of collapsing features onto a single text embedding. To prevent task-specific fine-tuning from destroying CLIP's generalizable structure, the **U**-shaped **D**ual **S**pace **A**daptation (**U-DSA**) module introduces deep yet parameter-efficient adaptation and remaps adapted features back to shallower, more domain-invariant layers. Extensive experiments on WMCA, CeFA, PADISI, and SURF under complete-modality, missing-modality, and limited-source protocols show that MMDA consistently surpasses state-of-the-art methods, reducing average HTER by up to 9.63% and improving AUC by up to 5.98% in cross-domain evaluation.

Keywords: Face anti-spoofing · Multimodal · Domain Generalization

† These authors contributed equally to this work.
⋆ Corresponding authors

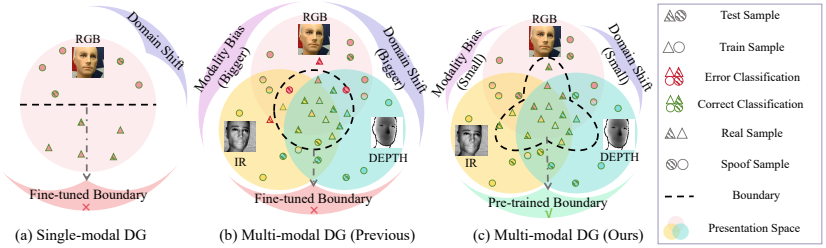

Fig. 1: (a) In single-modal FAS, domain shifts between training and testing environments induce domain gaps that degrade generalization. (b) In multimodal FAS, existing methods typically fine-tune a backbone and classifier on all source domains to learn a new decision boundary in the joint multimodal space; this boundary is often overly smoothed to accommodate diverse domain–modality combinations, causing hard target samples to lie near the boundary. (c) Our MMDA framework instead first reduces domain and modality gaps in the representation and then aligns multimodal features to a well-structured real/spoof separation inherited from the pre-trained space, improving separability under severe shifts. **Note:** Face images are from the WMCA [11] dataset.

1 Introduction

Facial recognition (FR) systems are widely deployed in high-stakes applications such as payment, access control, and surveillance [44, 53], and thus require reliable face anti-spoofing (FAS) [4, 16, 50]. Modern FAS systems increasingly adopt multimodal sensing with RGB, infrared (IR), and DEPTH streams, since multimodal inputs provide complementary cues on appearance, geometry, and material properties that are crucial for distinguishing real faces from spoof artifacts. However, multimodal FAS still generalizes poorly to unseen domains: performance often drops sharply under new cameras, environments, and attack types.

A distinctive difficulty of multimodal cross-domain FAS (beyond single modality DG FAS) is that generalization is challenged by two coupled shifts. First, the fused representation can drift across groups because cross-modal relations (e.g., complementarity, reliability, and scale among RGB/IR/DEPTH) are themselves domain dependent. Second, learning a new decision boundary by fine-tuning on limited source domains often yields an overly smoothed separator in the joint multimodal space, leaving hard target samples close to the boundary (Fig. 1(b)). To better understand the representation drift, we conduct two simple diagnostics on per-class embeddings for each modality (and the fused representation) and observe consistent cross-domain discrepancies. (i) An ANOVA-style ratio $r = \mathrm{Var}_{\mathrm{between}}/\mathrm{Var}_{\mathrm{total}}$ indicates that cross-domain variation, especially for bonafide (real) samples, exhibits a strong mean-shift component, suggesting the presence of an approximately additive bias. (ii) Same-domain pairs are significantly closer than cross-domain pairs under both ℓ_2 and cosine distances with separated bootstrap confidence intervals, implying that local neighborhoods are relatively stable within a group. Taken together, these statistics suggest that a non-trivial portion of the cross-group discrepancy behaves like a group-shared residual that can be estimated under same-group conditioning and removed before any alignment.

These observations motivate our *purify first, then guide* design. Instead of directly aligning noisy multimodal features, where spoof-relevant cues are entangled with domain/modality artifacts, we first *purify* the fused embedding by suppressing group-dependent components, and then *guide* the purified features into a stable real/spoof geometry that resists boundary drift. To this end, we view a fused feature as the sum of invariant spoof cues and variant artifacts. To operationalize this view at the fusion level, we adopt a denoising-style attention strategy and reformulate differential attention for the multimodal DG setting. Conventional Differential Attention (DA) [46] applies two attention branches to views of the same sample and subtracts their responses to suppress sample-wise spurious patterns. In our **M**odality-**D**omain Joint **D**ifferential **A**ttention (**MD2A**) module, each sample keeps its own multimodal features in the main attention branch, while the noise branch is fed with features from another sample in the same domain. The two branches thus share domain conditions and modality configuration but differ in identity and attack content. The differential response between them predominantly captures domain- and modality-specific components rather than semantics that distinguish real from spoof. Treating this response as a residual and subtracting it from the main branch output can help reduce shared domain-/modality-related variations, allowing the fused multimodal representation to better emphasize spoof-discriminative cues.

Once the features are purified, the next question is how to align them without destroying their structure. Aligning to a target defined only by limited source-domain data risks encoding the same biases we aim to remove, and hard alignment of visual features to a single text embedding can collapse the rich spoof manifold. To avoid such destructive alignment, we leverage a pre-trained CLIP model and its joint visual–text space as a strong, domain-agnostic anchor for real/spoof separation. We encode multiple textual descriptions of real and spoof faces into a text-defined subspace and introduce a **R**epresentation **S**pace **S**oft (**RS2**) alignment scheme that softly pulls multimodal visual features toward appropriate regions of this subspace. RS2 imposes a soft semantic constraint that guiding features toward the correct semantic neighborhood instead of pinning them to a single prompt, so that the model can exploit CLIP's semantic structure without losing fine-grained spoof cues.

Finally, we address a deeper tension in using a pre-trained model for FAS: without adapting deep layers, the feature extractor lacks task-specific capacity; yet aggressively fine-tuning deep layers tends to distort the pre-trained representation space and harm the generalization of its decision boundary. To balance these effects, we design a **U**-shaped **D**ual **S**pace **A**daptation (**U-DSA**) module. U-DSA introduces a deep, parameter-efficient adaptation pathway on top of CLIP while keeping the original backbone space as a stable anchor. The adapted deep features are then mapped back to shallower, more domain-invariant layers through a U-shaped structure, and RS2 is applied at multiple depths. This builds a bridge between a task-specialized space and the pre-trained space, enabling the model to increase task capacity while preserving the generalizable

CLIP geometry after MD2A has reduced domain and modality gaps. To sum up, our contributions are five-fold:

- We formulate multimodal DG FAS as a "purify then guide" problem and propose MMDA, a CLIP-based framework that first purifies multimodal features and then performs soft, non-destructive alignment.
- We introduce MD2A, which reformulates differential attention by feeding same-domain different-sample features into the noise branch, explicitly attenuating domain- and modality-specific artifacts in multimodal fusion while preserving discriminative spoof information. Moreover, we theoretically show that the proposed MD2A can tighten the generalization error bound in multimodal FAS scenarios.
- We develop RS2 and U-DSA, which together perform global and layer-wise soft alignment to a CLIP-based real/spoof representation space, balancing task-specific adaptation and preservation of pre-trained geometry.
- We conduct extensive experiments on four multimodal FAS benchmarks under complete-modality, missing-modality, and limited-source protocols, achieving state-of-the-art cross-domain performance.

2 Related Work

2.1 Face Anti-Spoofing

Deep learning has driven rapid progress in face anti-spoofing (FAS) [36, 39, 41], with numerous architectures designed to extract discriminative cues for separating real from spoof faces. However, performance typically degrades severely under domain shifts (e.g., changes in lighting, cameras, or acquisition setups) [13, 20, 21, 47, 53]. To improve domain generalization (DG), recent works explore adversarial learning [15,54], feature disentanglement [16,31,58], meta-learning [2,6], data augmentation [3, 9], and domain alignment [12, 19, 38, 40], aiming to learn domain-invariant feature representations and more robust decision boundaries. These DG-oriented methods are almost exclusively developed for unimodal FAS and do not account for multimodal fusion or modality-specific biases.

Multimodal FAS integrates multiple sensing streams to leverage complementary appearance, geometric, and material cues [17, 18, 22, 24, 28, 32, 33, 43, 52, 55]. Attention-based fusion and adaptive loss functions have been used to extract complementary information [10, 57], while cross-modal translation is explored to reduce semantic discrepancies across modalities [27]. Recent works further study robustness under missing-modality conditions by proposing new protocols and methods for flexible multimodal FAS [48, 49, 51]. To support arbitrary modality combinations, cross-modal attention and multimodal adapters with pre-trained ViTs are used to learn modality-insensitive features [23, 25, 28]. Among multimodal FAS methods that explicitly consider DG, MMDG [23] is a representative work: it analyzes the conflict between DG objectives and multimodal fusion, and introduces uncertainty rectification and modality re-balancing to alleviate this issue. In this paper, we follow this line of work and further investigate

representation-space methods that jointly handle domain and modality gaps for multimodal DG FAS using a CLIP-based framework.

2.2 CLIP-based FAS

Parameter-efficient transfer learning adapts large pre-trained models such as Vision Transformers (ViT) [5] and CLIP [34] to new tasks by tuning a small subset of parameters, improving efficiency and reducing overfitting. PETL-style designs have shown strong performance in FAS [1, 4, 37]. For example, S-Adapter [4] inserts lightweight adapter modules into pre-trained backbones to adjust features in a parameter-efficient way, and rehearsal-based PETL frameworks [1] and FLIP-like designs [37] further demonstrate the benefit of adapting only a small number of parameters for robust face anti-spoofing performance.

Leveraging CLIP as a backbone, several works explore CLIP-based and vision–language FAS models [7, 29, 30, 37], typically using text prompts to provide semantic guidance and improve generalization [8]. These methods, however, are mostly unimodal or rely on simple late fusion, and primarily focus on how to update or regularize pre-trained weights and prompts. In multimodal DG FAS, the structure of the pre-trained representation space itself may offer further potential to be exploited or preserved, which motivates exploring representation-space-oriented designs on top of CLIP.

3 Method

Our MMDA framework is illustrated in Fig. 2(a). Following the "purify then guide" principle, we first reduce domain and modality artifacts in multimodal visual embeddings, and then softly align the purified features in a CLIP-based representation space. Concretely, input RGB/IR/DEPTH images and text prompts are fed into the CLIP image and text encoders to obtain visual and textual embeddings. Throughout training, we keep the CLIP text encoder frozen and fine-tune the CLIP image encoder on the FAS data. The resulting multimodal visual embeddings are first fused and denoised by Modality-Domain Joint Differential Attention (MD2A), and are then passed through the proposed U-shaped Dual Space Adaptation (U-DSA) module and aligned to text-defined real/spoof subspaces via the Representation Space Soft (RS2) scheme. Finally, a shallow classifier operates jointly on the adapted visual embeddings and text embeddings for real/spoof prediction.

3.1 Preliminary and Motivation

CLIP for FAS and a hidden risk. A common CLIP-based FAS recipe [21,37] is to construct text prompts describing real/spoof faces and classify an image by its similarity to these prompts, optionally combined with parameter-efficient tuning of the visual branch. However, the CLIP pre-training objective does not explicitly enforce cross-domain or cross-modality consistency for FAS. In multimodal DG FAS, directly fine-tuning the visual encoder and learning a

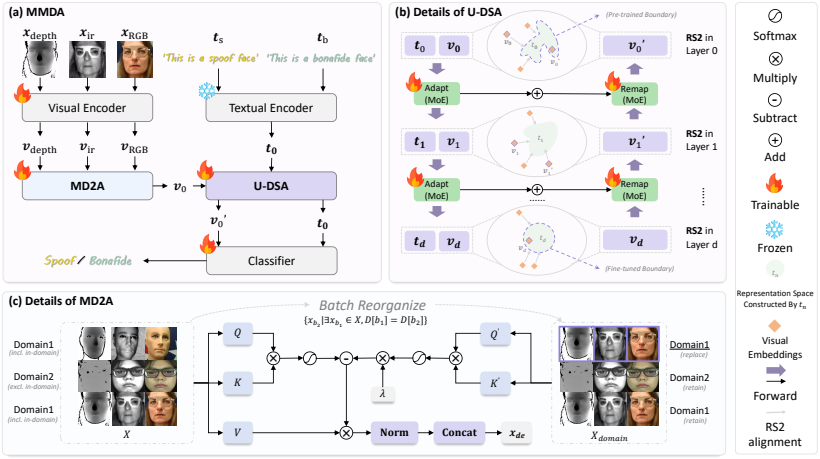

Fig. 2: Overall framework of MMDA. (a) Following the "purify then guide" principle, MMDA first denoises multimodal visual embeddings with Modality-Domain Joint Differential Attention (MD2A), and then performs soft, non-destructive alignment in a CLIP-based representation space. (b) Details of the U-shaped Dual Space Adaptation (U-DSA) module and the layer-wise use of the Representation Space Soft (RS2) alignment scheme, which together balance task-specific adaptation and preservation of CLIP geometry. (c) Operational details of MD2A, where cross-sample same-domain pairing is used to estimate domain/modality-shared residuals. **Note:** All face images shown are from the WMCA [11], CeFA [26], PADISI [35], and SURF [56] datasets.

new decision boundary from limited source domains can therefore (i) overfit to acquisition-specific artifacts and (ii) distort the pre-trained geometry, i.e., the decision boundary drifts away from a more transferable separator induced by the frozen text space. This tension motivates using CLIP as an anchor while constraining task adaptation to be non-destructive.

Groups: domain and modality setting. Each multimodal sample is collected under a domain $d \in \mathcal{D}$ (dataset/camera/environment) and a modality setting $m \in \mathcal{M}$ (available sensing streams, e.g., RGB+IR+DEPTH or missing IR/DEPTH). We denote a group by $g = (d, m) \in \mathcal{G} := \mathcal{D} \times \mathcal{M}$. Different groups induce different acquisition biases and fusion behaviors, so multimodal embeddings can shift systematically across g.

A decomposition view and the "purify then guide" principle. Let $h(x)$ denote the fused multimodal embedding before our purification module. Motivated by the empirical mean-shift evidence, we adopt a simple decomposition view:

$$h(x) = s(x, y) + \Delta_g + \varepsilon, \tag{1}$$

where $s(x, y)$ captures spoof-relevant, group-invariant semantics and Δ_g is a group-dependent bias term. This suggests two design requirements for multimodal DG FAS: (i) *Purify*: suppress Δ_g before any alignment, so that cross-group discrepancies (the "shift term" in our analysis) are reduced. (ii) *Guide*

Algorithm 1: Modality–Domain Joint Differential Attention (MD2A)

Input: batch multimodal features $X = \{\boldsymbol{x}_i\}_{i=1}^{B}$; domain labels $\mathcal{D} = \{\boldsymbol{d}_i\}_{i=1}^{B}$
Output: purified multimodal features X_{de}

```
# 1. Construct same-domain pairs (preferably different samples)
for i ← 1 to B do
    choose j s.t. D[j] = D[i] (and j ≠ i if possible);
    x̃_i ← Concat(x_i, x_j);
end
X̃ ← {x̃_i}_{i=1}^B;
# 2. Project and split into main / differential branches
Q_all ← X̃W_q,   K_all ← X̃W_k,   V ← X̃W_v;
(Q, Q') ← split(Q_all),   (K, K') ← split(K_all);
# 3. Differential attention for feature purification
s ← 1/√n_d;
A ← QK^⊤ · s,   A' ← Q'K'^⊤ · s;
X_de ← (softmax(A) − λ softmax(A')) V;
```

without destroying geometry: align purified features to a stable semantic structure while preventing boundary drift, which corresponds to constraining the learned separator to stay close to a pre-trained one.

Concretely, our MD2A operationalizes (i) by using same-group conditioning (in practice, same-domain pairing; and same modality setting when applicable) to estimate group-shared residual patterns and subtract them, producing a purified embedding that is easier to align robustly. Based on purified features, RS2 and U-DSA address (ii): RS2 provides a soft semantic pull by aligning to class-specific text subspaces spanned by multiple prompts, and U-DSA increases task capacity via deep adaptation while remapping back to shallower, more domain-invariant spaces to preserve CLIP geometry.

3.2 Modality-Domain Joint Differential Attention

MD2A implements the *purify* step in our framework by explicitly estimating and suppressing domain- and modality-specific residuals in multimodal features. As shown in Fig. 2(c), let $X = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_b\}$ denote fused multimodal features in a mini-batch and $\mathcal{D}$ the corresponding domain labels. For each sample $\boldsymbol{x}_i$, we randomly select another sample $\tilde{\boldsymbol{x}}_i$ from the same domain:

$$\mathcal{D}[\tilde{\boldsymbol{x}}_i] = \mathcal{D}[\boldsymbol{x}_i], \quad \tilde{\boldsymbol{x}}_i \neq \boldsymbol{x}_i \text{ if available}, \tag{2}$$

and fall back to $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i$ when no other same-domain sample exists. In practice, we implement this pairing by reorganizing the batch as in Alg. 1, where each row of X is replaced by the feature-wise concatenation of $(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i)$ using `Concat`, which is applied in both training and inference.

We then project the concatenated features and split them into a main branch and a differential branch:

$$(Q, Q') = \mathrm{split}(XW_q), (K, K') = \mathrm{split}(XW_k), V = XW_v, \tag{3}$$

where W_q, W_k, W_v are learnable projection matrices and n_d is the feature dimension. The attention maps of the two branches are

$$A = \frac{QK^\top}{\sqrt{n_d}}, \quad A' = \frac{Q'K'^\top}{\sqrt{n_d}}. \tag{4}$$

The differential branch learns a group-shared residual bias associated with Δ_g and suppresses it in the fused features. The MD2A output is obtained by contrasting the two branches:

$$X_{de} = \left[\mathrm{softmax}(A) - \lambda\,\mathrm{softmax}(A')\right] V, \tag{5}$$

where λ controls the strength of the differential branch, and we follow the original differential attention [46] for its parameterization.

$$\lambda = \exp(\lambda_{q1} \cdot \lambda_{k1}) - \exp(\lambda_{q2} \cdot \lambda_{k2}) + 0.8 - 0.6 \times \exp(-0.3 \cdot (l-1)), \tag{6}$$

where $\lambda_{q1}, \lambda_{k1}, \lambda_{q2}, \lambda_{k2} \in \mathbb{R}^{n_d}$ are learnable vectors, and $l \in [1, L]$ denotes the layer index. In practice, we set $l = 14$. When $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i$ for all i, MD2A degenerates to the original differential attention, which suppresses sample-wise noise. By feeding same-domain (preferably different) samples into the differential branch, MD2A estimates a group-shared residual bias associated with Δ_g. Subtracting this residual attenuates domain- and modality-specific artifacts while preserving sample-specific, spoof-relevant cues in the fused representation, providing a cleaner starting point for subsequent alignment.

3.3 Representation Space Alignment

Representation Space Soft Alignment (RS2). RS2 realizes the "guide" step by softly pulling purified multimodal features toward appropriate regions of a text-defined real/spoof subspace in the CLIP representation space, rather than collapsing them onto a single prompt. As shown in Fig. 2(b), given a set of captions C, we obtain text embeddings T via the frozen CLIP text encoder. Let V denote the corresponding visual embeddings produced by the CLIP image encoder, MD2A and U-DSA. RS2 encourages each visual embedding to lie close to the subspace spanned by text embeddings of the same class, while a shared classifier provides additional discriminative supervision on both modalities.

Let $\hat{y}_i \in \{0, 1\}$ be the ground-truth label (real/spoof) of a visual embedding $\boldsymbol{v}_i$, and $T^{(\hat{y}_i)} \subset T$ the set of text embeddings for class $\hat{y}_i$. We define its distance to the class-specific text subspace as:

$$d_i = \min_{\boldsymbol{t}_j \in T^{(\hat{y}_i)}} \frac{1}{2}\left(1 - \frac{\boldsymbol{v}_i \cdot \boldsymbol{t}_j}{\|\boldsymbol{v}_i\|\,\|\boldsymbol{t}_j\|}\right), \tag{7}$$

i.e., the minimum cosine distance to text embeddings of the same class. To avoid over-confident targets, we apply label smoothing: for a $\epsilon = 0.2$:

$$\tilde{y}_i = \begin{cases} 1 - \epsilon, & \hat{y}_i = 1, \\ \epsilon, & \hat{y}_i = 0. \end{cases} \tag{8}$$

The alignment loss is a smoothed binary cross-entropy on the distance-based score:

$$\mathcal{L}_{\text{align}} = -\sum_{\boldsymbol{v}_i \in V} \left(\tilde{y}_i \log(1 - d_i) + (1 - \tilde{y}_i) \log d_i\right). \tag{9}$$

We further attach a shared binary classifier on top of both visual and text embeddings. Let $\boldsymbol{e}_j$ denote either a visual or a text embedding, and $p_{\boldsymbol{e}_j}$ the predicted probability of the real class. Using the same smoothed targets, the classification loss is:

$$\mathcal{L}_{\text{cls}} = -\sum_{\boldsymbol{e}_j \in \{V,T\}} \left(\tilde{y}_j \log p_{\boldsymbol{e}_j} + (1 - \tilde{y}_j) \log(1 - p_{\boldsymbol{e}_j})\right). \tag{10}$$

The RS2 loss combines both terms:

$$\mathcal{L}_{\text{RS2}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{align}}. \tag{11}$$

Compared to hard alignment to a single text prototype, RS2 provides a soft semantic pull toward a *subspace* defined by multiple prompts, allowing visual features to exploit CLIP's semantic structure while retaining fine-grained spoof cues.

U-shaped Dual Space Adaptation (U-DSA). RS2 relies on a meaningful CLIP geometry, but task-specific adaptation on limited FAS data can easily distort this geometry and harm cross-domain generalization. Simply stacking deeper adaptation layers boosts capacity at the cost of warping the pre-trained space. U-DSA is designed to resolve this tension by (i) applying RS2 layer-wise and (ii) building a U-shaped dual-space pathway that feeds deep adapted features back to shallower, more domain-invariant spaces.

Let $\boldsymbol{v}_0$ be the input visual embedding from the CLIP image encoder (after MD2A), and let d be the depth of U-DSA. The forward (bottom) path progressively applies lightweight adaptation modules:

$$\boldsymbol{v}_i = \text{Adapt}_i(\boldsymbol{v}_{i-1}), \quad i = 1, \ldots, d, \tag{12}$$

where each Adapt_i is implemented by a parameter-efficient MoE. The feedback (top) path then remaps deeper representations back to shallower spaces:

$$\boldsymbol{v}'_d = \boldsymbol{v}_d, \quad \boldsymbol{v}'_i = \boldsymbol{v}_i + \text{Remap}_i(\boldsymbol{v}'_{i+1}), \quad i = d-1, \ldots, 0, \tag{13}$$

where Remap_i (also implemented as MoE) projects the deeper representation $\boldsymbol{v}'_{i+1}$ into the space of layer i. In practice, RS2 losses are applied to intermediate representations $\boldsymbol{v}_i$ and $\boldsymbol{v}'_i$ to regularize each layer toward the text-defined subspaces.

This U-shaped dual-space design allows deeper layers to capture task-specific multimodal patterns, while the feedback path and layer-wise RS2 keep the overall representation close to the pre-trained CLIP geometry. As a result, the final classifier operates in a space that is both enriched by adaptation and anchored by the original real/spoof separation, completing the "purify then guide" pipeline after MD2A has removed most domain and modality artifacts.

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 [14]	26.09	82.03	28.50	75.91	41.82	60.56	40.48	62.31	37.32	68.25
SSAN [42]	17.73	91.69	27.94	79.04	34.49	68.85	36.43	69.29	35.34	70.98
SA-FAS [38]	21.37	87.65	23.22	84.49	35.10	70.86	35.38	69.71	28.77	78.18
IADG [58]	27.02	86.50	23.04	83.11	32.06	73.83	39.24	63.68	39.83	62.95
FLIP [25]	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	36.60	68.12
AMA [49]	17.56	88.74	27.50	80.00	21.18	85.51	47.48	55.56	27.47	79.85
VP-FAS [48]	16.26	91.22	24.42	81.07	21.76	85.46	39.35	66.55	29.82	76.62
ViTAF [13]	20.58	85.82	29.16	77.80	30.75	73.03	39.75	63.44	33.89	71.54
MM-CDCN [52]	38.92	65.39	42.93	59.79	41.38	61.51	48.14	53.71	46.81	53.43
CMFL [10]	18.22	88.82	31.20	75.66	26.68	80.85	36.93	66.82	31.01	75.07
MMDG [23]	12.79	93.83	15.32	92.86	18.95	88.64	29.93	76.52	19.24	87.96
DADM [45]	11.71	94.89	6.92	97.66	19.03	88.22	16.87	91.08	13.63	92.96
CLIP [34]	14.55	90.47	18.17	90.02	24.13	83.15	38.33	65.71	24.63	83.00
MMDA (Ours)	**1.22**	**99.99**	**4.21**	**98.62**	**4.34**	**98.58**	**6.25**	**98.18**	**4.00**	**98.94**

4 Experiments

4.1 Experimental Setup

Datasets and protocols. We follow the multimodal DG protocol of MMDG [23] and evaluate MMDA on four benchmarks: WMCA (**W**) [11], CeFA (**C**) [26], PADISI (**P**) [35], and SURF (**S**) [56]. Three types of protocols are considered: (i) *Protocol 1* (complete modalities) uses all available modalities for training and testing under four Leave-One-Out (LOO) domain splits, e.g., **CPS→W** trains on **C**, **P**, **S**, and tests on **W**; (ii) *Protocol 2* introduces test-time missing-modality conditions on top of Protocol 1, including missing DEPTH, missing IR, and missing both DEPTH and IR; (iii) *Protocol 3* uses only two datasets as source domains (CW→PS and PS→CW) to simulate limited-source training.

Metrics. Following prior work [23, 48, 49], we report Half Total Error Rate (HTER) and Area Under the ROC Curve (AUC) for all protocols. Lower HTER and higher AUC indicate better performance.

Implementation details. We adopt CLIP ViT-B/16 [34] as the backbone. The CLIP text encoder is kept frozen, while the CLIP image encoder is fine-tuned jointly with the proposed MD2A and U-DSA modules and the final classifier on the multimodal FAS data. All images are resized to 224×224 and tokenized into 14×14 patch tokens with a 512-dim embedding. We train MMDA with AdamW, a learning rate of 5×10^{-6}, weight decay 1×10^{-3}, batch size 576, and 100 epochs for all protocols. Unless otherwise specified, U-DSA uses 7 layers; the classifier operates on the adapted visual embeddings.

4.2 Cross-domain Testing

Protocol 1: Complete modality scenario. Protocol 1 evaluates cross-domain generalization when all modalities are available in both training and testing. As shown in Table 1, MMDA consistently outperforms both unimodal DG baselines and multimodal FAS methods on all four LOO splits. Compared with the

Table 2: Cross-dataset testing results under the missing modalities scenarios (Protocol 2) among CASIA-CeFA (C), PADISI (P), CASIA-SURF (S), and WMCA (W). 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 [14]	38.92	65.45	37.64	66.57	39.18	65.22	38.58	65.75
SSAN [42]	36.77	69.21	41.20	61.92	33.52	73.38	37.16	68.17
SA-FAS [38]	36.30	69.07	39.80	62.69	33.08	74.29	36.40	68.68
IADG [58]	40.72	58.72	42.17	61.83	37.50	66.90	40.13	62.49
FLIP [25]	23.66	83.90	24.06	84.04	27.07	79.79	27.93	79.44
Multi-modal FAS								
ViT [5]	40.04	64.69	36.77	68.19	36.20	69.02	37.67	67.30
AMA [49]	29.25	77.70	32.30	74.06	31.48	75.82	31.01	75.86
VP-FAS [48]	29.13	78.27	29.63	77.51	30.47	76.31	29.74	77.36
ViTAF [13]	34.99	73.22	35.88	69.40	35.89	69.61	35.59	70.64
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 [23]	24.89	82.39	23.39	83.82	25.26	81.86	24.51	82.69
DADM [45]	21.56	85.17	20.82	85.28	22.61	84.04	21.66	84.83
CLIP [34]	28.07	77.00	29.10	77.04	32.58	73.36	33.83	71.11
MMDA (Ours)	**11.10**	**93.97**	**5.98**	**98.30**	**13.36**	**93.74**	**10.14**	**95.33**

strongest prior multimodal DG method DADM [45], MMDA reduces the average HTER from 13.63% to 4.00% and increases the average AUC from 92.96% to 98.94% (+5.98%). All four sub-protocols achieve HTER $\leq$ 6.25% and AUC $\geq$ 98.18%, indicating that the learned multimodal representations transfer well to unseen domains. These results support our design choice of first shrinking domain and modality gaps via MD2A and then aligning multimodal features in a CLIP-based representation space using RS2 and U-DSA, rather than relearning a new fused decision boundary from scratch.

Protocol 2: Missing modality scenario at test time. Protocol 2 evaluates robustness when some modalities are missing during testing. Table 2 shows that MMDA again achieves the best performance in all three missing-modality settings. On average, MMDA reduces HTER from 21.66% (DADM) to 10.14% and improves AUC from 84.83% to 95.33%, without any modality-specific dropout or retraining for missing modalities. When IR is missing, MMDA still obtains 5.98% HTER and 98.30% AUC, very close to the complete-modality case, suggesting that the purified RGB and DEPTH features after MD2A and their soft alignment via RS2 can compensate for the loss of IR. In contrast, removing DEPTH leads to a larger degradation (11.10% HTER and 93.97% AUC), highlighting that DEPTH provides complementary geometric information that is harder to recover from RGB and IR alone, even under the proposed purify then guide scheme.

Protocol 3: Limited source domain scenario. Protocol 3 restricts the number of source domains to two, simulating limited-source training. In Table 3, MMDA performs strongly in both CW→PS and PS→CW. In the CW→PS setting, MMDA improves HTER from 12.61% (DADM) to 7.52% and AUC from 93.81% to 96.84%. In the more challenging PS→CW setting, where WMCA shows substantial distribution shift, MMDA reduces HTER from 20.40% to 6.30% and raises AUC from 89.51% to 98.35%. The improvements under limited-

Table 3: Cross-dataset testing results under the limited source domain scenarios (Protocol 3) among CASIA-CeFA (C), PADISI (P), CASIA-SURF (S), and WMCA (W). The best results are in **bold**.

Method	CW→PS		PS→CW	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
Uni-modal DG (Concat + 1×1 Conv)				
SSDG [14]	25.34	80.17	46.98	54.29
SSAN [42]	26.55	80.06	39.10	67.19
SA-FAS [38]	25.20	81.06	36.59	70.03
IADG [58]	22.82	83.85	39.70	63.46
FLIP [25]	15.92	92.38	23.85	83.46
Multi-modal FAS				
ViT [5]	42.66	57.80	42.75	60.41
AMA [49]	29.25	76.89	38.06	67.64
VP-FAS [48]	25.90	81.79	44.37	60.83
ViTAF [13]	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 [23]	20.12	88.24	36.60	70.35
DADM [45]	12.61	93.81	20.40	89.51
CLIP [34]	19.36	90.57	29.98	79.22
MMDA (Ours)	**7.52**	**96.84**	**6.30**	**98.35**

Table 4: Ablation results on MD2A. The best results are in **bold**.

Method	HTER(%)↓	AUC(%)↑
Dense Adaptor	23.26	84.92
Dense Adaptor (w/ MHSA)	25.85	82.95
Dense Adaptor (w/ DA)	16.49	92.05
Dense Adaptor (w/ MD2A)	**13.47**	**94.20**
MoE Adaptor	22.92	85.84
MoE Adaptor (w/ MHSA)	12.83	93.25
MoE Adaptor (w/ DA)	12.72	93.89
MoE Adaptor (w/ MD2A)	**9.70**	**95.23**

Table 5: Ablation results on RS2. The best results are in **bold**.

Method	HTER(%)↓	AUC(%)↑
Vanilla Alignment	9.70	95.23
Smooth Alignment	9.17	96.32
RS2 Alignment	**8.88**	**97.20**

Table 6: Pairing-swap invariance analysis on MD2A.

Pairing strategy	Var[R] ↓	Var[z] ↓
Same-domain	4.31×10^{-7}	1.672×10^{-3}
Random	7.48×10^{-7}	1.662×10^{-3}
Cross-domain	6.94×10^{-7}	1.728×10^{-3}

source training further confirm that explicitly purifying multimodal features and leveraging CLIP-based soft alignment allows MMDA to retain generalizable cues even when domain coverage is scarce.

4.3 Ablation Study

Effectiveness of MD2A. Table 4 evaluates MD2A under two variants on the PS→CW split. Starting from the Dense adapter baseline (23.26% HTER / 84.92% AUC), adding a standard multi-head self-attention (MHSA) even hurts performance. Replacing MHSA with Differential Attention (DA) [46] yields clear gains (16.49% / 92.05%), and our MD2A further reduces HTER to 13.47% and raises AUC to 94.20%. A similar trend holds for the MoE adapter: MD2A improves over both MHSA and DA, achieving 9.70% HTER and 95.23% AUC. These results indicate that reformulating DA with same-domain, cross-sample pairing indeed helps shrink domain and modality gaps in the fused representation, going beyond sample-wise DA and better realizing the "purify" step in our framework.

Effectiveness of RS2 Alignment. Table 5 studies different alignment strategies in the CLIP representation space. "Vanilla Alignment" directly penalizes distances between visual features and their corresponding text prompts, obtaining 9.70% HTER and 95.23% AUC. Introducing "Smooth Alignment" improves both metrics (9.17% / 96.32%), suggesting that overly hard alignment targets are suboptimal. Our full RS2 loss, which combines smoothed supervision with nearest-text subspace alignment and a shared classifier on both visual and text embeddings, further reduces HTER to 8.88% and increases AUC to 97.20%. This confirms that explicitly shaping the geometry of the CLIP space through soft subspace alignment, rather than merely updating model weights or aligning fea-

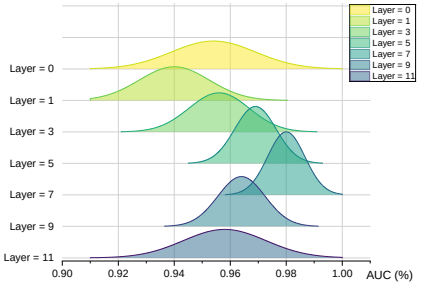

Fig. 3: AUC statistics of the U-DSA module across various caption groups at different depths. The height of each bar represents the number of captions achieving the specified AUC.

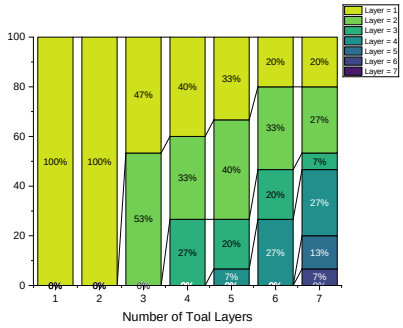

Fig. 4: Statistics of the layers achieving the best alignment effects in representation spaces constructed by U-DSA under different total depths (1 to 7 layers).

tures with a single prompt, yields consistent improvements in multimodal DG FAS and effectively realizes the "guide" step.

Effectiveness of U-DSA. We next analyze the impact of U-DSA depth and its dual-space design. Fig. 3 summarizes, for different caption sets (representation spaces), how many caption groups achieve a given AUC at each U-DSA depth. The distribution shifts towards higher AUC values as the depth increases up to 7 layers, indicating that deeper adaptation modules help capture richer multimodal patterns when combined with MD2A and RS2. A complementary view is given in Fig. 4: for total depths from 1 to 7, the layers that yield the best performance are predominantly shallow or mid-depth, and the deepest layer never achieves the best AUC. This matches our design intuition for U-DSA: deep adaptation is useful, but the final classifier should operate on a remapped representation closer to the shallower CLIP space, rather than on the deepest features alone.

Denoising mechanism analysis. We interpret MD2A as estimating and cancelling domain-/modality-correlated common-mode nuisance under same-domain conditioning, thereby purifying the representation while preserving spoof discriminative cues. To validate this mechanism, we propose a pairing-swap invariance test (Table 6). Specifically, for each sample x, the main branch outputs $F_{\text{main}}(x)$. Given a paired sample $\tilde{x}$, the differential branch predicts a residual $R(x, \tilde{x})$, and we form the denoised feature as

$$F_{\text{de}}(x, \tilde{x}) = F_{\text{main}}(x) - R(x, \tilde{x}). \tag{14}$$

We keep x fixed and only swap $\tilde{x}$ with three pairing strategies (*same-domain* / *random* / *cross-domain*), repeating pairing 10 times for each x. We then measure: (i) **residual stability** via within-swap $\text{Var}[R]$; (ii) **decision stability** via within-swap $\text{Var}[z]$, where z is the spoof logit computed from F_{de}; and (iii) **domain leakage** via HSIC between features and domain labels (lower indicates weaker dependence). As shown in Table 6, *same-domain* pairing yields the lowest $\text{Var}[R]$ and exhibits a clear mean-residual drift when switching to *cross-domain* pairing (drift magnitude: $9.55{\times}10^{-7}$). Meanwhile, domain dependence drops sub-

Fig. 5: t-SNE visualization of the fine-tuned CLIP baseline (left) and the classifier part of MMDA (right).

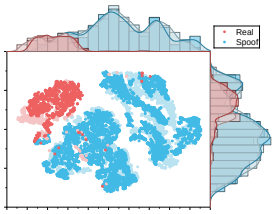

Fig. 6: t-SNE visualization of representations before entering U-DSA (lighter color) and after the deepest layer (darker color).

stantially after subtraction (HSIC: $1.596 \rightarrow 0.269$, drop $= 1.327$), while $\mathrm{Var}[z]$ remains small. Together, these results support that MD2A performs common-mode noise suppression rather than arbitrary semantic subtraction, without destabilizing spoof decisions.

4.4 Visualization and Analysis

The t-SNE plots in Fig. 5 compare the fine-tuned CLIP baseline with MMDA. For CLIP, real and spoof samples from different domains form several overlapping clusters, and cross-domain mixing within each class is limited. In contrast, MMDA produces a much clearer separation between real and spoof, with samples from different domains better mixed inside each class cluster, consistent with the goal of first reducing domain and modality gaps and then aligning to a stable CLIP-based real/spoof separation.

Fig. 6 visualizes how representations evolve along U-DSA. From the input (lighter points) to the output after deep adaptation and remapping (darker points), samples gradually concentrate into tighter, more compact clusters. This indicates that U-DSA refines features while keeping them close to a stable decision structure, rather than drifting away from the pre-trained CLIP geometry.

5 Conclusion

We presented **M**ultimodal **D**enoising and **A**lignment (MMDA), a CLIP-based framework that tackles multimodal domain-generalized face anti-spoofing from a "purify then guide" perspective: MD2A first *purifies* fused RGB/IR/DEPTH features by suppressing domain-/modality-specific artifacts, and RS2 together with U-DSA then *softly guides* the purified representations within a CLIP-based real/spoof space without destroying its geometry. Extensive experiments on WMCA, CeFA, PADISI, and SURF under complete-modality, missing-modality, and limited-source protocols show substantial and consistent gains over prior unimodal DG and multimodal FAS methods. Limitations remain, e.g., mixture-of-experts layers can be optimization-sensitive in pre-trained spaces. Future work will focus on more stable and adaptive dual-space adaptation and richer, more inclusive caption sets to further enhance generalization across unseen domains, sensors, and attack types.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62576216 and 62576076), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037), the Guangdong Provincial Key Laboratory (Grant No. 2023B1212060076), the CCF–Tencent Rhino-Bird Open Research Fund, and the Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047). Computational resources were provided by the SongShan Lake HPC Center (SSL-HPC) at Great Bay University.

References

1. Cai, R., Cui, Y., Li, Z., Yu, Z., Li, H., Hu, Y., Kot, A.: Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. In: Proceedings of the IEEE/CVF ICCV. pp. 8037–8048 (2023)
2. 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)
3. 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)
4. Cai, R., Yu, Z., Kong, C., Li, H., Chen, C., Hu, Y., Kot, A.C.: S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens. IEEE TIFS (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. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: CLIP-Adapter: better vision-language models with feature adapters. IJCV **132**(2), 581–595 (2024)
9. 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)
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. 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)
13. 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)

14. 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)
15. 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)
16. 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)
17. 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)
18. 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)
19. 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)
20. 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)
21. 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)
22. 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 (2025)
23. 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)
24. Lin, X., Wang, S., Yu, Y., Yu, Z., Zhou, J., Liu, Y., Cao, X., Kot, A., Zheng, Y.: Learning representation and synergy invariances: A povable framework for generalized multimodal face anti-spoofing. arXiv preprint arXiv:2511.14157 (2025)
25. Liu, A., Liang, Y.: Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. arXiv:2304.07549 (2023)
26. 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)
27. 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)
28. 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)
29. 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)
30. Liu, S.Q., Wang, Q., Yuen, P.C.: Bottom-up domain prompt tuning for generalized face anti-spoofing. In: ECCV. pp. 170–187. Springer (2025)
31. 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)

32. Ma, H., Liu, A., Lin, X., Escalante, H.J., Guyon, I., Wan, J., Lei, Z., Liang, Y.: Cm-clip: Few-shot cross-modal generalization for face anti-spoofing via prompt learning. Machine Intelligence Research **23**(2), 366–382 (2026)
33. 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)
34. 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)
35. 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)
36. 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)
37. Srivatsan, K., Naseer, M., Nandakumar, K.: FLIP: cross-domain face anti-spoofing with language guidance. In: Proceedings of the IEEE/CVF ICCV. pp. 19685–19696 (2023)
38. 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)
39. Wang, H., Shi, Y., Tao, Z., Gao, Y., Zhang, L., Lin, X., Feng, J., Yuan, X., Yu, Z., Cao, X.: Faceshield: Explainable face anti-spoofing with multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 9811–9819 (2026)
40. 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)
41. Wang, T., Lin, X., Xu, Y., Ye, Q., Guo, D., Escalera, S., Khoriba, G., Yu, Z.: Micro-gesture recognition: A comprehensive survey of datasets, methods, and challenges. Machine Intelligence Research **23**(2), 308–330 (2026)
42. 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)
43. 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)
44. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE TPAMI **45**(10), 12113–12132 (2023)
45. 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)
46. Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., Wei, F.: Differential transformer. ICLR (2025)
47. 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)
48. Yu, Z., Cai, R., Cui, Y., Liu, A., Chen, C.: Visual prompt flexible-modal face anti-spoofing. TDSC (2024)

49. 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)
50. Yu, Z., Cai, R., Li, Z., Yang, W., Shi, J., Kot, A.C.: Benchmarking joint face spoofing and forgery detection with visual and physiological cues. IEEE TDSC (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, J., Lin, X., Huang, J., Ye, S., Guo, X., Zhu, D., Hu, R., Guo, D., Liang, Y., Yu, Z., et al.: Multimodal deception detection: A survey. Machine Intelligence Research **23**(2), 284–307 (2026)
56. 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)
57. 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)
58. 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)