Single-Pass Evidence Measurement for Interpretable and Uncertainty-Aware 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

Face anti-spoofing (FAS) is critical for protecting face recognition systems against presentation attacks in open environments. Beyond high cross-domain generalization, practical FAS deployments must provide interpretable and uncertainty-aware decisions when multimodal evidence is incomplete, degraded, or conflicting. However, current multimodal domain-generalized FAS methods focus primarily on feature fusion or domain alignment, compressing RGB, depth, and infrared cues into opaque logits. This makes it difficult to discern which modality drives the decision, whether modalities reinforce or suppress each other, and whether a given sample lies within the classifier's reliable operating regime. We introduce **SEM-FAS**, a quantum-inspired yet fully classical *single-pass evidence measurement* framework that jointly enhances generalization, interpretability, and reliability. SEM-FAS encodes each modality as a normalized complex-valued state whose amplitude captures evidence strength and whose phase encodes cross-modal interference; multiple coherent heads are then mixed into a density-matrix representation. A constrained POVM-like measurement head converts this representation into calibrated class probabilities while analytically decomposing every prediction into per-modality main effects and pairwise synergy/suppression terms. Simultaneously, three built-in uncertainty indicators are obtained without sampling overhead: predictive ambiguity, representation mixedness, and measurement-coverage mismatch. On standard multimodal DG-FAS benchmarks with complete modalities, SEM-FAS outperforms the strongest baseline by 2.18 percentage points in average HTER and 1.39 percentage points in average AUC. Visual analyses further suggest that the learned modality effects, pairwise interactions, and uncertainty scores align with expected FAS behavior, even for hard samples, indicating that single-pass evidence measurement is a promising way to jointly enhance generalization, interpretability, and reliability in multimodal FAS.

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

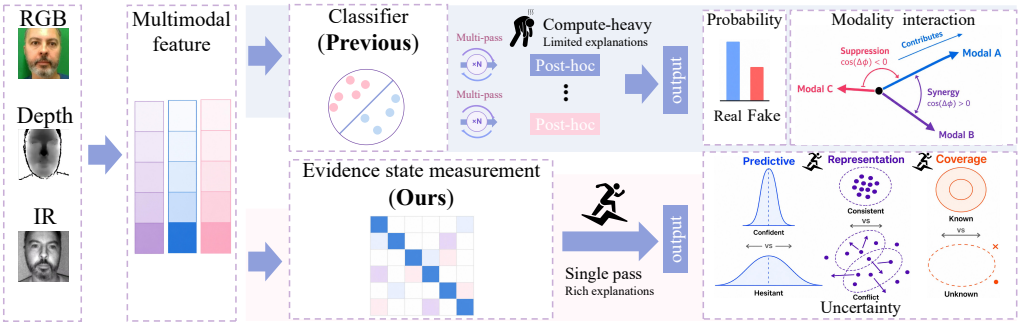

Figure 1: **Classical vs. Measurement-Based Modeling.** Classical multimodal classifiers mainly output final probabilities, while richer information such as modality interactions, predictive confidence, representation consistency, and coverage is not directly available. Recovering these signals usually requires additional post-hoc modules, often with multiple inference passes and heavy computation. In contrast, our method obtains classification results and rich explanations in a single forward pass, enabling efficient uncertainty-aware and interaction-aware decision-making. **Note:** Face images are from the WMCA (8) dataset.

1 Introduction

With the widespread deployment of face recognition in security-critical applications (e.g., access control, smartphone unlocking, mobile payment), face anti-spoofing (FAS) has become a vital front-end module for detecting presentation attacks such as printed photos, replayed videos, and 3D masks. Although recent deep FAS models (39; 16) have achieved strong performance on standard benchmarks, their reliability often degrades under unseen sensors, illumination conditions, capture environments, and attack materials. Two complementary directions have emerged to address this challenge: domain-generalized FAS, which aims to generalize to unseen target domains without accessing target-domain labels, and multimodal FAS, which exploits RGB, depth, and infrared inputs to provide complementary texture, geometric, and spectral cues.

Along the first direction, domain generalization (DG) techniques have been introduced into FAS (14; 45) to alleviate domain shift. Along the second, multimodal data compensate for the limitations of single-modality decisions under challenging illumination or high-fidelity attacks (42; 38; 31; 36). More recent studies (20; 37; 26; 19; 24; 25) unify both directions within joint frameworks, improving the generalization and stability of multimodal FAS on unseen domains.

However, robust classification alone is insufficient for trustworthy multimodal FAS deployment. We argue that multimodal DG-FAS should also be viewed as a structured evidence measurement problem. In practice, low confidence can arise from fundamentally different causes. It may arise from *information insufficiency* caused by domain shift, imaging degradation, missing modalities, or low-quality acquisition; it may also arise from *evidence conflict*, where different modalities support inconsistent live/spoof hypotheses. For example, a printed attack may preserve plausible RGB texture while producing abnormal depth or infrared responses, whereas low-quality acquisition may weaken all modalities without a clear spoofing cue. These cases require different downstream actions, such as rejection, recapture, or human inspection. However, a standard softmax posterior only provides a scalar confidence score and cannot explicitly answer which modality supports the decision, whether modalities reinforce or suppress each other, or whether the sample lies within the reliable coverage of the learned classifier. Existing multimodal DG-FAS methods mainly address robustness at the representation level, while their final decision heads remain standard logit-based classifiers whose scores are not decomposable into modality-wise evidence or pairwise interactions. While post-hoc tools such as Shapley-style attribution (27) or Monte Carlo dropout (6) can partially fill this gap, they require repeated forward passes and are expensive for latency-sensitive FAS deployment.

To address this, we propose **SEM-FAS**, a *single-pass evidence measurement* framework for trustworthy multimodal DG-FAS. The central idea is to replace the unconstrained logit head with a structured measurement head that preserves multimodal evidence throughout the final decision. We adopt quantum state-measurement formalism as a compact mathematical scaffold (quantum-inspired but fully classical) because it naturally pairs a structured state with a constrained measurement that yields decomposable probabilities. Specifically, each modality is mapped to a normalized complex

*Equal contribution.
†Corresponding authors.

state whose **amplitude** controls evidence allocation and whose **phase** enables pairwise interference, constructive when modalities agree, destructive when they conflict. Multiple coherent heads form a **density-matrix-style representation** that preserves coexisting evidence patterns. A constrained POVM-like measurement head produces class probabilities via quadratic scores while analytically decomposing every prediction into modality main effects and pairwise synergy/suppression terms. An additional **reject effect** provides a coverage signal; together with predictive entropy and state purity, this yields three endogenous uncertainty indicators, label ambiguity, representation mixedness, and measurement coverage mismatch, all in a single forward pass without post-hoc sampling, as illustrated in Fig. 1.

This measurement view is particularly aligned with the practical challenges of multimodal FAS. Constructive interactions capture cases where RGB, depth, and infrared jointly support the same live/spoof hypothesis, while suppressive interactions indicate cross-modal inconsistency. The mixed evidence state represents coexisting evidence patterns across coherent heads, which is useful when a sample contains both bona fide-like and attack-like cues. The additional reject measurement channel provides a coverage signal: when a sample is weakly supported by the learned task measurements, more energy is assigned to the reject effect.

Our contributions are summarized as follows:

- We formulate multimodal DG-FAS as a structured evidence measurement problem, requiring a system to jointly produce classification, modality-wise evidence, cross-modal interaction, and uncertainty/coverage signals.
- We introduce a single-pass measurement head that maps modality features into a mixed-state representation and performs constrained quadratic measurement, enabling score-level decomposition into modality main effects and pairwise synergy/suppression terms.
- We derive three endogenous uncertainty signals from the same measurement process, predictive entropy, representation mixedness, and measurement coverage/reject energy, requiring no repeated stochastic passes or perturbations.
- Extensive experiments show that SEM-FAS improves multimodal DG-FAS under complete and missing-modality protocols, while yielding stronger calibration and more efficient uncertainty ranking than deterministic or post-hoc uncertainty baselines.

2 Related Work

2.1 Multimodal Domain Generalization FAS

Domain generalization (DG) for single-modal FAS (1; 40; 17; 18) has attracted considerable attention, with typical approaches using adversarial learning (12; 30) or style augmentation (45) to reduce domain shift, and meta-learning schemes (47; 5) that simulate train/test splits to improve adaptation to new domains. In parallel, multimodal FAS exploits complementary appearance, spectral, and geometric cues from RGB, depth, and IR, and recent methods based on cross-modal attention (21; 23) or multimodal adapters (42) show that learning modality-invariant features in a unified pretrained space can improve cross-domain robustness. However, directly transferring single-modal DG strategies to multimodal FAS has been reported to yield limited or even negative gains. Multimodal DG-FAS therefore aims to jointly exploit multimodal cues and generalize across domains within a single framework. MMDG (20) attributes performance bottlenecks to the coupling between modality bias and domain shift and introduces uncertainty-aware modality rebalancing, while DADM (37) adopts a dual-alignment scheme to correct inter- and intra-domain modality misalignment. Despite these advances, existing methods largely retain a standard binary classification head: fusion and alignment are handled at the feature level, but the final score is not decomposable into per-modality evidence or pairwise interactions. Even when uncertainty is modeled, as in MMDG, it serves as an internal reweighting signal rather than a deployment-facing explanation.

2.2 Quantum-inspired Deep Learning Methods

Quantum-inspired (QI) approaches (10; 35) bring the formalism of quantum states and measurements into classical models to impose structured inductive biases. Existing work is broadly state-centric or measurement-centric. State-centric methods map embeddings to Hilbert-space representations to

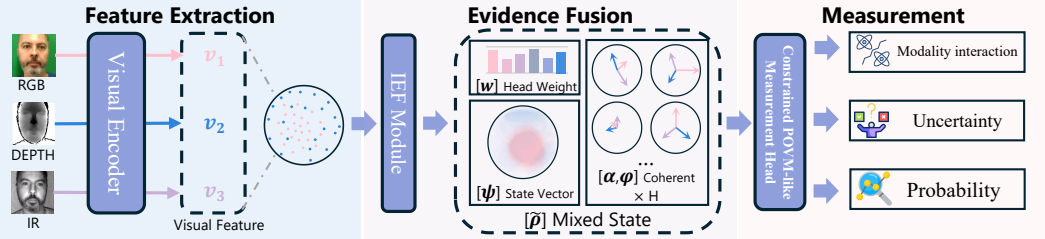

Figure 2: Overview of SEM-FAS. A shared visual encoder extracts real-valued features from multimodal inputs (RGB/DEPTH/IR). The interference-based evidence fusion (IEF) module maps them to modality-level pure states and constructs a mixed state $\tilde{\rho}$. A constrained POVM-like measurement head then performs measurement on $\tilde{\rho}$ to obtain class probabilities and interpretable indicators, including modality interaction, and multi-level uncertainty. **Note:** Face images are from the WMCA (8) dataset.

model mixed evidence and richer geometry, for example via quantum-inspired projection heads for similarity learning (13). Measurement-centric methods use learned measurements for probabilistic decisions or structured transformations, including state-discrimination classifiers (3), measurement-like updates for multimodal sequence fusion (15), and generalized measurements implemented as learnable POVM modules (34) or mixed-state self-attention (2). These studies demonstrate that state and measurement structures can serve as useful inductive biases for classification or representation learning. However, none of them leverage the formalism to make the final classification score itself analytically decomposable into per-modality evidence contributions and pairwise interaction terms.

2.3 Interpretability and Uncertainty Estimation

Interpretability and uncertainty estimation are essential for trustworthy multimodal FAS, as practitioners need to understand which modalities drive a prediction and how confident the system is before acting on it. Existing tools are predominantly post-hoc. Perturbation-based attribution methods such as Shapley-style approaches (27) require repeated forward passes to assign credit to each modality, while Monte Carlo dropout (6) estimates uncertainty through stochastic sampling at inference time. Both incur significant computational overhead impractical for latency-sensitive FAS deployment. More fundamentally, they provide neither *score-decompositional* explanations, where modality main effects and pairwise interactions are analytically derived from the classification score itself, nor *structurally separated* uncertainty indicators that distinguish label ambiguity, representation-level mixture, and measurement coverage mismatch. SEM-FAS addresses these gaps by making interpretability and uncertainty *endogenous* to the classification process: the measurement score is analytically decomposable into per-modality and pairwise terms, and three distinct uncertainty signals emerge from a single forward pass without additional sampling or perturbation.

3 Method

3.1 Preliminaries and Overview

Our objective is to design a single-pass decision head that jointly produces class probabilities, score-level evidence decomposition, and uncertainty signals for multimodal FAS. Instead of using an unconstrained logit classifier after feature fusion, SEM-FAS treats the final prediction as a structured evidence measurement process. The design borrows the quantum state-measurement formalism as a structured parameterization for evidence allocation, cross-modal interaction, and measurement coverage, while operating entirely in the classical regime.

In the quantum-inspired formalism, a state is represented by a density operator $\rho \in \mathbb{C}^{D\times D}$ ($\rho = \rho^\dagger$, $\rho \succeq 0$, $\mathrm{Tr}(\rho) = 1$), and observation is modeled by a POVM $\{E_j\}_{j=1}^{J}$ with $E_j \succeq 0$, $\sum_j E_j = I$, yielding outcome probabilities $s_j = \mathrm{Tr}(E_j\rho)$ via the Born rule. For classification, E_j acts as a class-related measurement effect and s_j as the projected evidence energy. Such scores are non-negative, comparable across effects, and naturally compatible with evidence decomposition and rejection (details in Sec. B).

As shown in Fig. 2, SEM-FAS contains three main components. Given a multimodal input $X = \{x_1, \ldots, x_M\}$ (e.g., RGB/IR/DEPTH), we first use a shared visual encoder $f(\cdot)$ to extract per-modality features $\mathbf{v}_m = f(x_m) \in \mathbb{R}^D$. Then, the Interference-based Evidence Fusion (IEF) module maps $\mathbf{v}_m$ to normalized complex modality states $|\psi_m\rangle$ and predicts multi-head amplitudes $\alpha_{h,m}$ and phases $\phi_{h,m}$. The amplitudes control modality-level evidence allocation, while phase differences induce pairwise synergy or suppression. The head-wise interference states are mixed with weights w_h to form a density-matrix-style mixed evidence state $\tilde{\rho}$ that preserves coexisting evidence patterns. Finally, a low-rank, approximately complete POVM-like measurement head with an additional reject channel performs a single measurement on $\tilde{\rho}$ to obtain scores $\{s_j\}$. The same constrained quadratic scores are used for classification, modality main effects, pairwise interaction terms, and uncertainty/coverage indicators on a common measurement scale.

3.2 Interference-based Evidence Fusion (IEF)

Complex state embedding. We model each modality as a normalized pure state in $\mathbb{C}^D$. Given real-valued features $\mathbf{v}_m \in \mathbb{R}^D$ from the shared encoder, we obtain:

$$\mathbf{z}_m^r = \mathbf{W}_r \mathbf{v}_m + b_r, \quad \mathbf{z}_m^i = \mathbf{W}_i \mathbf{v}_m + b_i, \quad \psi_m = \frac{\mathbf{z}_m^r + i\,\mathbf{z}_m^i}{\sqrt{\|\mathbf{z}_m^r\|_2^2 + \|\mathbf{z}_m^i\|_2^2}}, \tag{1}$$

where $\mathbf{W}_r, \mathbf{W}_i \in \mathbb{R}^{D\times D}$ and $b_r, b_i \in \mathbb{R}^D$ are learnable. The resulting $\psi_m \in \mathbb{C}^D$ satisfies $\langle \psi_m, \psi_m \rangle = 1$ and is regarded as the modality pure state $|\psi_m\rangle$. A sample is represented as $\Psi = \{|\psi_1\rangle, \ldots, |\psi_M\rangle\}$.

Amplitude and phase generation. We introduce a lightweight coherence controller that predicts, for each head h and modality m, a non-negative amplitude $\alpha_{h,m}$ (evidence allocation) and a phase $\phi_{h,m} \in [-\pi, \pi]$ (interaction control). The amplitude controls how much evidence from modality m is allocated to head h, while the phase controls how this modality interacts with the others within the same head:

$$\alpha_{h,m} = \mathrm{softmax}\left(\frac{\mathbf{v}_m^\top \mathbf{q}_h^{(1)}}{\sqrt{D}}\right), \qquad \phi_{h,m} = \pi\,\tanh\Big((\mathbf{q}_h^{(2)})^\top \mathbf{v}_m\Big), \tag{2}$$

where $\mathbf{q}_h^{(1)}, \mathbf{q}_h^{(2)} \in \mathbb{R}^D$ are learnable query vectors of head h, and the softmax is over modalities.

Mixed-state preparation. The head-level state is defined by the coherent superposition $|u_h\rangle = \sum_{m=1}^M \sqrt{\alpha_{h,m}}\, e^{i\phi_{h,m}} |\psi_m\rangle$, $h = 1, \ldots, H$. To weight the heads per sample, we compute

$$\bar{\mathbf{v}} = \frac{1}{M}\sum_{m=1}^M \mathbf{v}_m, \qquad w_h = \mathrm{softmax}\Big((\mathbf{q}_h^{(3)})^\top \bar{\mathbf{v}}\Big), \tag{3}$$

where $\{\mathbf{q}_h^{(3)}\}_{h=1}^H$ are learnable and the softmax is over heads. Here $\alpha_{h,m}$ controls the energy allocation of modality m within head h; the phase difference $\phi_{h,m} - \phi_{h,n}$ determines constructive or destructive interaction between modality pair (m, n); and w_h modulates the contribution of different interference patterns at the sample level. The mixed evidence state is

$$\tilde{\rho} = \frac{\sum_{h=1}^H w_h\, |u_h\rangle\langle u_h|}{\mathrm{Tr}\Big(\sum_{h=1}^H w_h\, |u_h\rangle\langle u_h|\Big)} \in \mathbb{C}^{D\times D}, \tag{4}$$

which retains multiple evidence patterns instead of collapsing all modalities into a single fused vector before classification.

3.3 Constrained POVM-like Measurement Head

The measurement head maps the mixed evidence state to class scores while preserving a decomposable quadratic structure. We implement a set of POVM-like measurement operators $\mathcal{E} = \{E_1, \ldots, E_C, E_R\}$, where $E_1, \ldots, E_C$ correspond to the C task classes and E_R is an additional reject effect used only for analysis. Each effect is parameterized in a low-rank positive semidefinite form,

$$E_j = L_j L_j^\dagger, \qquad L_j = L_j^{(r)} + i L_j^{(i)}, \qquad j \in \{1, \ldots, C, R\}, \tag{5}$$

where $L_j^{(r)}, L_j^{(i)} \in \mathbb{R}^{D\times r}$ and $r \ll D$. This factorization guarantees $E_j \succeq 0$. To obtain an approximately complete measurement and avoid unbounded quadratic scores, we apply global spectral normalization to the raw effects so that their sum is constrained to be close to the identity, i.e., $\sum_{j=1}^{C} E_j + E_R \approx I$. The reject effect E_R absorbs residual measurement energy and is not used as a supervised class during training.

During forward inference, we do not explicitly construct $\tilde{\rho}$, but apply L_j directly to the head-level states. Let $Z = \mathrm{Tr}(\sum_h w_h |u_h\rangle\langle u_h|)$ be the normalization term in Eq. 4. The Born-type score is

$$s_j = \mathrm{Tr}(E_j\tilde{\rho}) = \frac{1}{Z}\sum_{h=1}^{H} w_h \|L_j^{\dagger} u_h\|_2^2, \qquad j \in \{1,\ldots,C,R\}. \tag{6}$$

Thus s_j is exactly the score obtained by measuring the normalized mixed state $\tilde{\rho}$, while implementation only requires weighted aggregation over the head-level states.

For supervised classification, we use only the first C class effects. The scores are mapped to class probabilities by

$$\ell_j = \log(s_j + \epsilon) + b_j, \qquad p(y{=}j \mid x) = \mathrm{softmax}_j(\ell_1,\ldots,\ell_C), \tag{7}$$

where b_j is a class bias and $\epsilon > 0$ ensures numerical stability. Since both logits and explanations derive from the same scores $\{s_j\}$, SEM-FAS provides score-level interpretability without a separate post-hoc model.

3.4 Evidence Decomposition and Uncertainty Indicators

The indicators in this subsection are not auxiliary probes added after classification. They are derived from the same state-measurement process used to compute the class scores in Eq. 6, and are therefore available in a single deterministic forward pass.

Multimodal evidence indicators. The measurement structure allows us to derive, in closed form, a modality main effect $F_{j,m}^{\mathrm{main}}$ (class-conditioned evidence allocated to modality m) and a pairwise interaction term $F_{j,mn}^{\mathrm{pair}}$ (signed cooperative/suppressive indicator) for each class j. The former measures how much of the class score is attributed to modality m, while the latter indicates whether modalities (m,n) cooperatively reinforce or suppressively weaken the class evidence. Because both indicators are defined with respect to the same normalized state $\tilde{\rho}$ and class effect E_j, they are tied to the classifier's own measurement score rather than to an external attribution objective. This places modality-level and pairwise contributions on a common measurement scale comparable across classes and samples. Definitions are in Appendix D.

Uncertainty indicators. The same $\tilde{\rho}$ and $\{E_j\}$ define three uncertainty indicators at different levels: (1) at the predictive level, the Shannon entropy $H(P)$ from the class probability vector measures label ambiguity; (2) at the representation level, state purity $\Pi(\tilde{\rho}) = \mathrm{Tr}(\tilde{\rho}^2)$ describes how mixed the evidence state is, where lower purity indicates that multiple evidence patterns are active and may correspond to heterogeneous or conflicting evidence; (3) at the coverage level, the reject score $s_R = \mathrm{Tr}(E_R\tilde{\rho})$ yields a task coverage score $\kappa = 1 - s_R$, where low κ signals that the input is weakly supported by the learned measurement effects. Together they describe predictive ambiguity, representation mixedness, and coverage mismatch. Unlike post-hoc attribution or Monte Carlo uncertainty estimation, these indicators are obtained in one deterministic forward pass from the same measurement process used for classification. Details are in Appendix E.

4 Experiments

4.1 Cross-Dataset Testing

Implementation Details. We adopt a pretrained CLIP ViT-B/16 encoder as the unified backbone. All modality images are resized to 224×224 and split into 14×14 patches; the resulting 768-D token embeddings are linearly projected to 512-D feature vectors. These per-modality vectors are fed into the IEF module with $H{=}8$ coherent heads and a constrained measurement head with low-rank factor $r{=}256$. We train end-to-end with AdamW (lr $1{\times}10^{-4}$, weight decay $1{\times}10^{-3}$, batch size 72, 100 epochs) and evaluate on four multimodal FAS benchmarks (WMCA (8), CeFA (22), PADISI (29),

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** and the second-best are underlined.

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 (11)	26.09	82.03	28.50	75.91	41.82	60.56	40.48	62.31	34.22	70.20
SSAN (33)	17.73	91.69	27.94	79.04	34.49	68.85	36.43	69.29	29.15	77.22
SA-FAS (32)	21.37	87.65	23.22	84.49	35.10	70.86	35.38	69.71	28.77	78.18
IADG (46)	27.02	86.50	23.04	83.11	32.06	73.83	39.24	63.68	30.34	76.78
CLIP (28)	14.55	90.47	18.17	90.02	24.13	83.15	38.33	65.71	23.80	82.34
FLIP (21)	13.19	93.79	11.73	94.93	17.39	90.63	22.14	83.95	16.11	90.83
Multi-modal FAS										
ViT (4)	20.88	84.77	44.05	57.94	33.58	71.80	42.15	56.45	35.17	67.74
AMA (42)	17.56	88.74	27.50	80.00	21.18	85.51	47.48	55.56	28.43	77.45
VP-FAS (41)	16.26	91.22	24.42	81.07	21.76	85.46	39.35	66.55	25.45	81.08
ViTAF (9)	20.58	85.82	29.16	77.80	30.75	73.03	39.75	63.44	30.06	75.02
MM-CDCN (43)	38.92	65.39	42.93	59.79	41.38	61.51	48.14	53.71	42.84	60.10
CMFL (7)	18.22	88.82	31.20	75.66	26.68	80.85	36.93	66.82	28.26	78.04
MMDG (20)	12.79	93.83	15.32	92.86	18.95	88.64	29.93	76.52	19.25	87.96
DADM (37)	11.71	94.89	**6.92**	**97.66**	19.03	88.22	16.87	91.08	13.63	92.96
SEM-FAS(Ours)	**4.32**	**99.21**	18.68	90.17	**15.91**	**90.74**	**6.89**	**97.27**	**11.45**	**94.35**

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** and the second-best are underlined.

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 (11)	38.92	65.45	37.64	66.57	39.18	65.22	38.58	65.75
SSAN (33)	36.77	69.21	41.20	61.92	33.52	73.38	37.16	68.17
SA-FAS (32)	36.30	69.07	39.80	62.69	33.08	74.29	36.39	68.68
IADG (46)	40.72	58.72	42.17	61.83	37.50	66.90	40.13	62.48
CLIP (28)	28.07	77.00	29.10	77.04	32.58	73.36	29.92	75.80
FLIP (21)	23.66	83.90	24.06	84.04	27.07	79.79	24.93	82.58
Multi-modal FAS								
ViT (4)	40.04	64.69	36.77	68.19	36.20	69.02	37.67	67.30
AMA (42)	29.25	77.70	32.30	74.06	31.48	75.82	31.01	75.86
VP-FAS (41)	29.13	78.27	29.63	77.51	30.47	76.31	29.74	77.36
ViTAF (9)	34.99	73.22	35.88	69.40	35.89	69.61	35.59	70.74
MM-CDCN (43)	44.90	55.35	43.60	58.38	44.54	55.08	44.35	56.27
CMFL (7)	31.37	74.62	30.55	75.42	31.89	74.29	31.27	74.78
MMDG (20)	24.89	82.39	23.39	83.82	25.26	81.86	24.51	82.69
DADM (37)	21.56	85.17	20.82	85.28	22.61	84.04	21.66	84.83
SEM-FAS(Ours)	**12.97**	**92.95**	**20.22**	**85.54**	**19.65**	**85.81**	**17.61**	**88.10**

SURF (44)) using HTER and AUC. Following MMDG (20), we adopt three protocols: (1) complete modalities, (2) missing modalities at test time, and (3) limited source domains (Appendix A).

Complete Modality Scenario. Table 1 shows Protocol 1 results where all modalities are available. SEM-FAS achieves the best HTER/AUC on three of four target domains, with notable gains on CPS→W (4.32%/99.21%) and PSW→C (6.89%/97.27%). Averaged over all settings, it reaches 11.45% HTER and 94.35% AUC, outperforming the second-best DADM by 2.18 pp in HTER and 1.39 pp in AUC—without any domain-alignment loss.

Missing Modality Scenario During Testing. Under Protocol 2 (Table 2), SEM-FAS improves the average from 21.66%/84.83% (DADM) to 17.61%/88.10%. Removing DEPTH causes modest loss (12.97%/92.95%), while removing IR or both non-RGB modalities results in larger drops, consistent with the evidence decomposition in Sec. 4.3. This asymmetric degradation confirms that SEM-FAS learns non-uniform modality reliance and remains discriminative when part of the evidence is absent.

Limited Source Domain Scenario. Under Protocol 3 (Table 3) with only two source domains, SEM-FAS achieves 13.01%/93.03% on CW→PS and 21.26%/85.75% on PS→CW, remaining close to DADM (gap <1 pp on CW→PS) despite not using explicit domain alignment. This is expected since SEM-FAS restructures only the decision interface; the competitive performance suggests it is complementary to representation-level domain alignment rather than a replacement under limited source coverage.

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** and the second-best are underlined.

Method	CW→PS		PS→CW	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
Uni-modal DG (Concat + 1*1 Conv)				
SSDG (11)	25.34	80.17	46.98	54.29
SSAN (33)	26.55	80.06	39.10	67.19
SA-FAS (32)	25.20	81.06	36.59	70.03
IADG (46)	22.82	83.85	39.70	63.46
CLIP (28)	19.36	90.57	29.98	79.22
FLIP (21)	15.92	92.38	23.85	83.46
Multi-modal FAS				
ViT (4)	42.66	57.80	42.75	60.41
AMA (42)	29.25	76.89	38.06	67.64
VP-FAS (41)	25.90	81.79	44.37	60.83
ViTAF (9)	29.64	77.36	39.93	61.31
MM-CDCN (43)	29.28	76.88	47.00	51.94
CMFL (7)	31.86	72.75	39.43	63.17
MMDG (20)	20.12	88.24	36.60	70.35
DADM (37)	**12.61**	**93.81**	**20.40**	**89.51**
SEM-FAS(Ours)	13.01	93.03	21.26	85.75

Table 4: Performance comparison between pure-state and mixed-state quantum-inspired state modeling. The best results are in **bold**.

State type	Measure rank	HTER(%)↓	AUC(%)↑
pure	64	14.36	91.87
pure	128	14.24	91.92
pure	256	16.14	91.83
pure	512	14.26	91.80
mixed	64	8.98	95.66
mixed	128	9.15	96.10
mixed	256	**8.53**	**96.27**
mixed	512	8.59	95.75

Table 5: Results of different numbers of coherence heads under mixed-state modeling. The best results are in **bold**.

State type	Heads number	HTER(%)↓	AUC(%)↑
mixed	8	4.80	**98.37**
mixed	16	8.26	96.26
mixed	32	9.09	95.57
mixed	64	8.53	96.27
mixed	128	**3.81**	98.32
mixed	256	10.17	95.21

Table 6: Ablation of the constrained POVM-like measurement head. We compare a linear baseline and the POVM-like measurement head (with and without constraints) under the same encoder and training setup. We report classification performance (HTER/AUC), AURC/AUROC for three endogenous uncertainty scores (entropy, state purity, and reject probability), and probabilistic calibration metrics (Brier score and negative log-likelihood, NLL). For the Linear baseline, only entropy-based uncertainty and calibration can be computed; purity- and reject-probability-based scores require the POVM-like measurement head and are thus omitted ("–").

Method	Head design		Cls. perf.		Entropy		Purity		Reject prob		Calibration	
	Quad	Phys. constr.	HTER(%)↓	AUC(%)↑	AURC(%)↓	AUROC(%)↑	AURC(%)↓	AUROC(%)↑	AURC(%)↓	AUROC(%)↑	Brier(%)↓	NLL↓
Linear baseline	×	×	11.99	92.67	5.02	79.29	–	–	–	–	23.40	55.42
POVM-like meas. head	✓	×	7.00	97.06	3.35	82.53	4.98	78.44	3.23	84.11	12.56	24.74
POVM-like meas. head	✓	✓	6.89	97.27	0.94	88.91	1.17	91.45	1.72	83.92	7.85	16.64

4.2 Ablation Study

Ablations use sub-protocols of Protocol 1: state modeling and coherent-head ablations (Tables 4, 5) on CPS→W; measurement-head ablation (Table 6) on PSW→C.

State Modeling. Table 4 shows the mixed-state variant consistently outperforms pure-state (e.g., 8.53%/96.27% vs. 14.24%/91.92% at r=256), confirming that preserving multiple coexisting evidence patterns through head-wise mixing provides a more flexible representation than collapsing all modalities into a single state. Among measurement ranks, r=256 yields the best trade-off.

Coherent Heads. Table 5 varies H under the mixed-state setting. Although H=128 achieves the lowest HTER (3.81%), H=8 attains the best AUC (98.37%) with far fewer parameters. This suggests that a small number of heads is sufficient to capture the principal coexisting evidence patterns; we adopt H=8 as the default.

Measurement Head. Table 6: the quadratic form yields the main classification gain (HTER 11.99%→7.00%); adding POVM constraints improves HTER/AUC only slightly (6.89%/97.27%) but substantially improves calibration (Brier 12.56→7.85, NLL 24.74→16.64) and uncertainty ranking (entropy AURC 3.35%→0.94%, purity AUROC 78.44%→91.45%). The quadratic form provides discriminative capacity; the constraints yield well-calibrated uncertainty indicators.

Comparison with Post-hoc Uncertainty Methods. Table 7: SEM-FAS adds minimal overhead over the quadratic head (0.54ms vs. 0.02ms, +0.8M params) while substantially improving calibration (Brier 12.56→7.85) and ranking (AUROC 84.11→91.45). Compared with MCD, it achieves better metrics at ∼70× lower latency (0.54ms vs. 38.14ms).

Table 7: Comparison of decision heads with post-hoc uncertainty quantification methods. Calibration: Brier/NLL; ranking: AURC/AUROC; overhead: **P(M)**=parameters in millions, **T(ms)**=per-sample head time, **Mem(MB)**=per-sample head memory. MCD denotes Monte Carlo Dropout.

Head+UQ	**P(M)**	**T(ms)↓**	**Mem(MB)↓**	**Brier%↓**	**NLL%↓**	**AURC%↓**	**AUROC%↑**
linear	149.6	0.01	470.70	23.40	55.42	5.02	79.29
linear+MCD	149.6	38.11	471.27	23.39	55.15	4.48	82.42
quadratic	150.1	0.02	470.78	12.56	24.74	3.23	84.11
quadratic+MCD	150.1	38.14	471.35	12.23	23.06	2.97	88.22
Ours	150.9	0.54	472.27	**7.85**	**16.64**	**0.94**	**91.45**

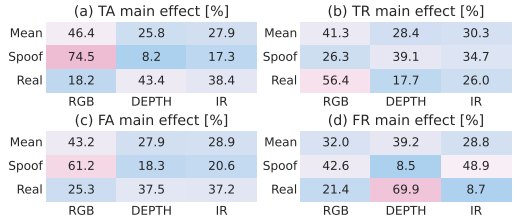

Figure 3: Visualization of modality main effects: normalized contributions of RGB, DEPTH, and IR to the Real/Spoof channels for the four decision types (TA/TR/FA/FR).

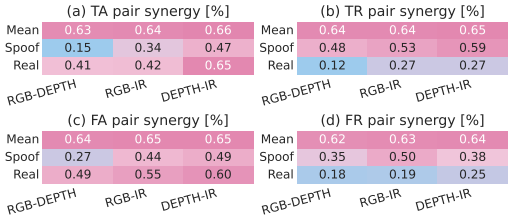

Figure 4: Visualization of modality pairwise effects: synergy coefficients between RGB, DEPTH, and IR in the Real/Spoof channels for the four decision types (TA/TR/FA/FR).

4.3 Visualization and Analysis

We analyze modality main effects and pairwise interactions at the dataset level, partitioning samples into TA, TR, FA, and FR, to verify whether the score-level decomposition from Sec. 3.4 is consistent with expected FAS behavior. Full distributions, head-level preferences, and case studies are in Appendix G.

Modality Main Effects. Figure 3 shows that for samples predicted as Real (TA/FA), the Real channel is mainly driven by DEPTH and IR while the residual Spoof response concentrates in RGB, indicating geometric/spectral cues provide bona fide evidence and RGB captures texture-related spoof cues. For Spoof predictions (TR/FR), abnormal DEPTH/IR patterns dominate the Spoof channel, and falsely rejected real faces often exhibit attack-like RGB/IR responses or degraded depth evidence. Every modality retains non-zero main effects across all outcomes, forming a "dominant plus auxiliary" pattern consistent with Protocol 2 results.

Modality Pairwise Effects. Figure 4 shows mostly cooperative pairwise interactions, with DEPTH–IR exhibiting slightly stronger synergy consistent with complementary geometric and spectral evidence. Synergy is systematically higher in the predicted class channel, indicating SEM-FAS amplifies cross-modal agreement; for failure cases (FA/FR), the wrong channel receives stronger support, helping diagnose errors.

5 Conclusion

We formulated multimodal domain-generalized FAS as a single-pass evidence measurement problem and proposed SEM-FAS, a quantum-inspired but fully classical framework that replaces the standard logit head with a constrained state-measurement interface. In one deterministic forward pass, SEM-FAS encodes multimodal features as density-matrix-style evidence states and performs POVM-like measurement, jointly producing classification probabilities, score-level modality main effects, pairwise synergy/suppression terms, and three endogenous uncertainty indicators, predictive entropy, representation mixedness, and measurement coverage, without post-hoc sampling or repeated inference. Experiments on four multimodal FAS benchmarks demonstrate that SEM-FAS achieves competitive or superior cross-domain generalization under complete, missing-modality, and limited source-domain protocols, while providing substantially stronger calibration and uncertainty ranking than both deterministic and post-hoc baselines at negligible additional latency. The score-level decomposition further reveals interpretable evidence patterns consistent with expected modality roles and observed robustness behavior.

Limitations and Future Work. SEM-FAS restructures only the decision head and does not explicitly optimize domain alignment at the representation level; under extremely limited source domains its gains over alignment-based methods diminish (Protocol 3). Integrating SEM-FAS with representation-level domain alignment is a natural next step to address this gap.

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] Cai, R., Cui, Y., Yu, Z., Lin, X., Chen, C., Kot, A.: Rehearsal-free and efficient continual learning for cross-domain face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

[2] Chen, F., Zhao, Q., Feng, L., Chen, C., Lin, Y., Lin, J.: Quantum mixed-state self-attention network. Neural Networks **185**, 107123 (2025)

[3] Cruzeiro, E.Z., De Mol, C., Massar, S., Pironio, S.: Quantum-inspired classification based on quantum state discrimination. Quantum Machine Intelligence **6**(2), 79 (2024)

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

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

[6] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)

[7] 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)

[8] 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)

[9] 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)

[10] Huynh, L., Hong, J., Mian, A., Suzuki, H., Wu, Y., Camtepe, S.: Quantum-inspired machine learning: a survey. arXiv preprint arXiv:2308.11269 (2023)

[11] 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)

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

[13] Kankeu, I., Fritsch, S.G., Schönhoff, G., Mounzer, E., Lukowicz, P., Kiefer-Emmanouilidis, M.: Quantum-inspired embeddings projection and similarity metrics for representation learning. arXiv preprint arXiv:2501.04591 (2025)

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

[15] Li, Q., Gkoumas, D., Sordoni, A., Nie, J.Y., Melucci, M.: Quantum-inspired neural network for conversational emotion recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 13270–13278 (2021)

[16] Li, S., Dutta, V., He, X., Matsumaru, T.: Deep learning based one-class detection system for fake faces generated by gan network. Sensors **22**(20), 7767 (2022)

[17] 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)

[18] 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)

[19] 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)

[20] 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 Computer Vision and Pattern Recognition. pp. 211–221 (2024)

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

[22] 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)

[23] 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 Transactions on Information Forensics and Security **18**, 4775–4786 (2023)

[24] 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)

[25] 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)

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

[27] Parcalabescu, L., Frank, A.: Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. arXiv preprint arXiv:2212.08158 (2022)

[28] 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)

[29] 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)

[30] Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10023–10031 (2019)

[31] 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)

[32] 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 CVPR. pp. 24563–24574 (2023)

[33] 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)

[34] Wei, J., He, Z., Chen, C., Deng, M., Situ, H.: Povm-based quantum self-attention neural network. In: 2023 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR). pp. 117–122. IEEE (2023)

[35] Wu, H., Zhang, J., Wang, L., Li, D., Kong, D., Han, Y.: A survey on quantum deep learning. The Journal of Supercomputing **81**(4), 564 (2025)

[36] 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)

[37] 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)

[38] Yang, J., Yu, Z., He, J., Ni, X., Zhang, L., Li, H., Cao, X.: G 2 v 2 former: Graph guided video vision transformer for face anti-spoofing. IEEE Transactions on Information Forensics and Security (2025)

[39] Yang, X., Luo, W., Bao, L., Gao, Y., Gong, D., Zheng, S., Li, Z., Liu, W.: Face anti-spoofing: Model matters, so does data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3507–3516 (2019)

[40] 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)

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

[42] Yu, Z., Cai, R., Cui, Y., Liu, X., Hu, Y., Kot, A.C.: Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. International Journal of Computer Vision **132**(11), 5217–5238 (2024)

[43] 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)

[44] 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)

[45] Zhou, Q., Zhang, K.Y., Yao, T., Lu, X., Ding, S., Ma, L.: Test-time domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 175–187 (2024)

[46] 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)

[47] Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Ding, S., Ma, L.: Adaptive mixture of experts learning for generalizable face anti-spoofing. In: Proceedings of the 30th ACM international conference on multimedia. pp. 6009–6018 (2022)

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 (20). 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.

A.3 Evaluation Metrics

We report both classification performance and reliability-oriented metrics. For all metrics, lower HTER/AURC/Brier/NLL and higher AUC/AUROC indicate better performance.

Classification metrics. Following the FAS literature, we use Half Total Error Rate (HTER) and Area Under the ROC Curve (AUC). Let FAR and FRR denote the false acceptance and false rejection rates at a chosen operating point; then

$$\mathrm{HTER} = \frac{\mathrm{FAR} + \mathrm{FRR}}{2}. \tag{8}$$

AUC summarizes the trade-off between true positive rate and false positive rate over all thresholds; we report it as a percentage.

Uncertainty metrics. To assess uncertainty estimation, we adopt the Area Under the Risk–Coverage curve (AURC) and the AUROC between an uncertainty score and the error indicator. Let u_i be the uncertainty score for sample i, $e_i \in \{0, 1\}$ indicate misclassification (1 = error), and $\mathcal{R}(c)$ denote the empirical risk when keeping the c fraction of samples with lowest uncertainty. AURC is defined as

$$\mathrm{AURC} = \int_0^1 \mathcal{R}(c)\, dc, \tag{9}$$

approximated in practice by a discrete sum over coverage levels. AUROC measures how well u_i ranks errors ahead of correct predictions:

$$\mathrm{AUROC} = \Pr\big(u_i > u_j \,\big|\, e_i = 1,\, e_j = 0\big). \tag{10}$$

In our experiments, u_i is instantiated by three endogenous scores produced by the SEM-FAS head: (i) *predictive entropy* $H(P_i) = -\sum_j p_{i,j} \log p_{i,j}$ over the class-probability vector P_i; (ii) *state purity* $\Pi(\rho_i) = \mathrm{tr}(\rho_i^2)$ of the induced quantum-inspired state; and (iii) *reject probability* $p_{R,i}$ from the dedicated reject channel.

Calibration metrics. We quantify probabilistic calibration using the Brier score and negative log-likelihood (NLL). Let y_i be the ground-truth label for sample i, represented as a one-hot vector, and P_i the predicted class-probability vector. The Brier score is

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} \big\| P_i - y_i \big\|_2^2, \tag{11}$$

and NLL corresponds to the average cross-entropy loss:

$$\mathrm{NLL} = -\frac{1}{N}\sum_{i=1}^{N}\log p_{i,y_i}, \tag{12}$$

where p_{i,y_i} denotes the predicted probability of the true class.

B Background on quantum-inspired states and POVMs

For completeness, we briefly summarize the quantum-inspired notions used in the main paper. In our finite-dimensional parameterization, the state is represented by a density-matrix-style operator ρ defined on a complex feature space $\mathbb{C}^D$, which satisfies

$$\rho = \rho^\dagger, \qquad \rho \succeq 0, \qquad \mathrm{Tr}(\rho) = 1. \tag{13}$$

If there exists a normalized state vector $|\psi\rangle \in \mathbb{C}^D$ such that

$$\rho = |\psi\rangle\langle\psi|, \tag{14}$$

then ρ is called a pure state. More general mixed states can be written as a convex combination of pure states,

$$\rho = \sum_k p_k\,|\psi_k\rangle\langle\psi_k|, \qquad p_k \geq 0,\ \sum_k p_k = 1. \tag{15}$$

A positive operator-valued measure (POVM) describes a measurement process through a collection of measurement operators $\{E_j\}_{j=1}^{J}$ that satisfy

$$E_j \succeq 0, \qquad \sum_{j=1}^{J} E_j = I_D. \tag{16}$$

When a measurement is applied to a state ρ, the probability of obtaining outcome j is given by the Born rule:

$$p(y=j) = \mathrm{Tr}(E_j\rho). \tag{17}$$

C Approximate POVM normalization

In the main paper we parameterize each measurement operator as $E_j = L_j L_j^\dagger$ with low-rank factors $L_j \in \mathbb{C}^{D\times r}$. Directly learning $\{L_j\}$ makes it difficult to enforce the completeness constraint $\sum_j E_j = I_D$. We therefore start from a set of raw real factors $L_{j,\mathrm{raw}}^{(r)}, L_{j,\mathrm{raw}}^{(i)} \in \mathbb{R}^{D\times r}$ and define

$$S = \sum_{j=1}^{J}\left(L_{j,\mathrm{raw}}^{(r)}L_{j,\mathrm{raw}}^{(r)\top} + L_{j,\mathrm{raw}}^{(i)}L_{j,\mathrm{raw}}^{(i)\top}\right) + \delta I_D, \tag{18}$$

where $\delta > 0$ is a numerical stability term. Let $S = Q\Lambda Q^\top$ be the spectral decomposition with Λ diagonal and positive. We set

$$\begin{aligned} S^{-\frac{1}{2}} &= Q\Lambda^{-\frac{1}{2}}Q^\top, \\ L_j^{(r)} &= S^{-\frac{1}{2}}L_{j,\mathrm{raw}}^{(r)}, \\ L_j^{(i)} &= S^{-\frac{1}{2}}L_{j,\mathrm{raw}}^{(i)}, \end{aligned} \tag{19}$$

and construct $L_j = L_j^{(r)} + iL_j^{(i)}$. It then follows that

$$\begin{aligned} \sum_{j=1}^{J} E_j &= \sum_{j=1}^{J}\left(L_j^{(r)}L_j^{(r)\top} + L_j^{(i)}L_j^{(i)\top}\right) \\ &= S^{-\frac{1}{2}}(S - \delta I_D)S^{-\frac{1}{2}} \\ &= I_D - \delta S^{-1}. \end{aligned} \tag{20}$$

When δ is sufficiently small and S^{-1} is well conditioned, $\sum_j E_j$ is close to I_D, so the resulting structure can be regarded as an approximately complete POVM. This normalization stabilizes training and prevents the measurement operators from drifting to degenerate configurations.

D Multimodal interpretability indicators: definitions and implementation

In this section we detail the definitions and derivations of the modality main effects and pairwise effects, and relate these analytical quantities to the tensors used in our implementation.

D.1 Decomposition of the Born score

Recall that the head-level state for head h is

$$|u_h\rangle = \sum_{m=1}^{M} \sqrt{\alpha_{h,m}}\, e^{i\phi_{h,m}} |\psi_m\rangle, \tag{21}$$

and the per-head score for class j is

$$s_{j,h} = \|L_j^\dagger u_h\|_2^2 = \langle u_h | E_j | u_h \rangle \geq 0, \qquad E_j = L_j L_j^\dagger. \tag{22}$$

Since the head-level states are generally not normalized, we use the same normalization factor as in the main paper:

$$Z = \mathrm{Tr}\left(\sum_{h=1}^{H} w_h |u_h\rangle\langle u_h| \right) = \sum_{h=1}^{H} w_h \|u_h\|_2^2. \tag{23}$$

The mixed-state Born score s_j is obtained by measuring the normalized mixed state:

$$s_j = \frac{1}{Z} \sum_{h=1}^{H} w_h s_{j,h}, \qquad s_{j,h} = \|L_j^\dagger u_h\|_2^2. \tag{24}$$

This is consistent with the normalized score used in Eq. (6) of the main paper.

Using the class-dependent response vectors

$$t_{j,m} = L_j^\dagger |\psi_m\rangle \in \mathbb{C}^r, \qquad e_{j,m} = \|t_{j,m}\|_2^2 = \langle \psi_m | E_j | \psi_m \rangle, \tag{25}$$

we can write

$$L_j^\dagger u_h = \sum_{m=1}^{M} \sqrt{\alpha_{h,m}}\, e^{i\phi_{h,m}} t_{j,m}. \tag{26}$$

Therefore,

$$\begin{aligned} s_{j,h} = \left\| L_j^\dagger u_h \right\|_2^2 &= \left\| \sum_{m=1}^{M} \sqrt{\alpha_{h,m}} e^{i\phi_{h,m}} t_{j,m} \right\|_2^2 \\ &= \sum_{m=1}^{M} \alpha_{h,m} e_{j,m} + \sum_{m \neq n} \sqrt{\alpha_{h,m}\alpha_{h,n}}\, \kappa_{j,h,mn}, \end{aligned} \tag{27}$$

where the cross-term coefficient is

$$\kappa_{j,h,mn} = 2\,\mathrm{Re}\big(e^{i(\phi_{h,m} - \phi_{h,n})} \langle t_{j,m}, t_{j,n} \rangle \big). \tag{28}$$

Substituting Eq. (27) into Eq. (24) yields

$$\begin{aligned} s_j &= \frac{1}{Z} \sum_{h=1}^{H} w_h s_{j,h} \\ &= \underbrace{\sum_{m=1}^{M} \left(\frac{1}{Z} \sum_{h=1}^{H} w_h \alpha_{h,m} \right) e_{j,m}}_{\text{single-modality (main-effect) term}} \\ &\quad + \underbrace{\frac{1}{Z} \sum_{m \neq n} \sum_{h=1}^{H} w_h \sqrt{\alpha_{h,m}\alpha_{h,n}}\, \kappa_{j,h,mn}}_{\text{cross-modality (pairwise) term}}. \end{aligned} \tag{29}$$

This gives an exact decomposition of the normalized Born score s_j into single-modality and cross-modality contributions.

D.2 Definition of main and pairwise effects

From Eq. (29), we define the normalized aggregated amplitude weight for modality m as

$$\nu_m = \frac{1}{Z} \sum_{h=1}^{H} w_h \alpha_{h,m}. \tag{30}$$

The single-modality contribution can then be written as $\sum_m \nu_m e_{j,m}$. We define the modality main effect as

$$E_{j,m}^{\text{main}} = \nu_m e_{j,m} = \left(\frac{1}{Z} \sum_{h=1}^{H} w_h \alpha_{h,m} \right) e_{j,m}, \tag{31}$$

so that

$$\sum_{m=1}^{M} E_{j,m}^{\text{main}} \tag{32}$$

exactly equals the single-modality part of the normalized Born score s_j in Eq. (29). This shows that $E_{j,m}^{\text{main}}$ is a faithful measure of how much modality m contributes to the class-j score under the joint action of all interference heads.

For the cross-modality term, we further examine $\kappa_{j,h,mn}$. Writing

$$\langle t_{j,m}, t_{j,n} \rangle = a_{j,mn} + i b_{j,mn}, \tag{33}$$

with $a_{j,mn} = \text{Re}\langle t_{j,m}, t_{j,n} \rangle$ and $b_{j,mn} = \text{Im}\langle t_{j,m}, t_{j,n} \rangle$, we have

$$\begin{aligned} \kappa_{j,h,mn} &= 2\,\text{Re}\big(e^{i(\phi_{h,m} - \phi_{h,n})}(a_{j,mn} + i b_{j,mn})\big) \\ &= 2\big(\cos(\phi_{h,m} - \phi_{h,n})\, a_{j,mn} - \sin(\phi_{h,m} - \phi_{h,n})\, b_{j,mn}\big). \end{aligned} \tag{34}$$

In practice we keep the dominant real part and factor out a class-dependent Gram term and a class-agnostic coherence term. Specifically, we define

$$\tilde{G}_{j,mn} = a_{j,mn} = \text{Re}\big(\langle t_{j,m}, t_{j,n} \rangle\big), \tag{35}$$

and the normalized class-agnostic coherence term

$$\kappa_{mn}^{(Z)} = \frac{1}{Z} \sum_{h=1}^{H} w_h\, 2\sqrt{\alpha_{h,m}\alpha_{h,n}}\, \cos(\phi_{h,m} - \phi_{h,n}), \tag{36}$$

which depends only on $\{\alpha_{h,m}, \phi_{h,m}, w_h\}$ and the normalization factor Z, and is shared across classes. Approximating

$$\frac{1}{Z} \sum_{h=1}^{H} w_h \sqrt{\alpha_{h,m}\alpha_{h,n}}\, \kappa_{j,h,mn} \approx \tilde{G}_{j,mn}\, \kappa_{mn}^{(Z)}, \tag{37}$$

the cross-modality contribution to s_j can be written as

$$E_{j,mn}^{\text{pair-energy}} = \tilde{G}_{j,mn}\, \kappa_{mn}^{(Z)}. \tag{38}$$

To obtain a dimensionless indicator that is easier to interpret and visualize, we normalize by the per-modality energies and define

$$G_{j,mn} = \frac{\tilde{G}_{j,mn}}{\sqrt{e_{j,m} e_{j,n}}} \in [-1, 1], \tag{39}$$

and

$$E_{j,mn}^{\text{pair}} = G_{j,mn}\, \kappa_{mn}^{(Z)}. \tag{40}$$

A positive $E_{j,mn}^{\text{pair}}$ indicates that the modality pair (m, n) has a synergistic effect for class j, whereas a negative value corresponds to an inhibitory effect. Under the above approximation, the normalized class score can be written as

$$s_j \approx \sum_m E_{j,m}^{\text{main}} + \sum_{m \neq n} E_{j,mn}^{\text{pair-energy}}, \tag{41}$$

so the main and pairwise effects together account for the full quadratic structure of the normalized Born score.

D.3 Mapping to implemented tensors

We now connect the analytical quantities to the tensors used in our implementation. Let t_r and t_i denote the real and imaginary parts of $t_{j,m}$. Then

$$\begin{aligned} \tilde{G}_{j,mn} &= t_{j,m}^{(r)\top} t_{j,n}^{(r)} + t_{j,m}^{(i)\top} t_{j,n}^{(i)}, \\ e_{j,m} &= \|t_{j,m}\|_2^2 = t_{j,m}^{(r)\top} t_{j,m}^{(r)} + t_{j,m}^{(i)\top} t_{j,m}^{(i)}. \end{aligned} \tag{42}$$

In code, we compute

$$\texttt{real_Gram_full} = t_r t_r^\top + t_i t_i^\top, \tag{43}$$

whose first C channels correspond to $\tilde{G}_{j,mn}$. The normalized Gram term

$$G_{j,mn} = \frac{\tilde{G}_{j,mn}}{\sqrt{e_{j,m} e_{j,n}}} \tag{44}$$

is implemented as `gram_cos`. The normalized coherence term $\kappa_{mn}^{(Z)}$ in Eq. (36) is implemented as

$$\kappa_{mn}^{(Z)} = \texttt{pair_coeff}_{mn}^{(Z)} = \frac{1}{Z} \sum_{h=1}^{H} w_h \, 2\sqrt{\alpha_{h,m}\alpha_{h,n}} \, \cos(\phi_{h,m} - \phi_{h,n}), \tag{45}$$

where $\alpha_{h,m}$ and $\phi_{h,m}$ are produced by the coherence controller, and Z is the same state-normalization factor used for the Born score.

Combining these factors, the tensor

$$\texttt{token_pair}_{j,mn} = \tilde{G}_{j,mn} \, \kappa_{mn}^{(Z)} \tag{46}$$

provides the energy-level pairwise contribution $E_{j,mn}^{\text{pair-energy}}$, while

$$\texttt{token_pair_norm}_{j,mn} = G_{j,mn} \, \kappa_{mn}^{(Z)} \tag{47}$$

corresponds exactly to the dimensionless indicator $E_{j,mn}^{\text{pair}}$ used in our visualizations.

E Uncertainty indicators: predictive, representation, and coverage

Beyond interpretability at the level of class scores, we also expose a set of *uncertainty indicators* that are fully endogenous to the POVM head, in the sense that they are computed directly from the same internal quantities used for prediction, without any auxiliary networks or calibration models. These indicators operate at three complementary levels:

- predictive uncertainty from the class-probability vector,
- representation-level uncertainty from the mixed state induced by the interference heads,
- coverage-based uncertainty from the POVM energy and reject channel.

Together, they provide a multi-view characterization of how confident the model is, from the output distribution back to the underlying quantum-like representation.

E.1 Predictive uncertainty from class probabilities

Let $P \in \mathbb{R}^C$ denote the class-probability vector after the POVM head and softmax. A standard measure of *predictive* uncertainty is the Shannon entropy

$$H(P) = -\sum_{j=1}^{C} p_j \log p_j, \tag{48}$$

which attains its minimum when a single class dominates and is maximized for a uniform distribution. In practice we also consider simple derived quantities such as the maximum class probability

$$p_{\max} = \max_j p_j \tag{49}$$

and the top-two probability margin

$$\mathrm{margin}(P) = p_{(1)} - p_{(2)}, \tag{50}$$

where $p_{(1)} \geq p_{(2)} \geq \cdots$ are the sorted probabilities. Low $p_{\max}$, small margin, and high entropy all indicate that the classifier is indecisive among several competing classes.

In the implementation, P is stored as `P`, and $p_{\max}$ and the margin are derived on the fly when constructing scalar uncertainty summaries and reject scores.

E.2 Representation-level uncertainty from state purity

At the level of internal representations, the interference heads induce a mixed state whose *purity* serves as a representation-level uncertainty indicator. Each head h produces a (generally unnormalized) state vector $|u_h\rangle \in \mathbb{C}^D$, and the heads are combined with nonnegative weights w_h satisfying $\sum_{h=1}^{H} w_h = 1$. We associate to this ensemble the (unnormalized) density operator

$$\tilde{\rho} = \sum_{h=1}^{H} w_h \, |u_h\rangle\langle u_h|, \tag{51}$$

whose trace

$$T = \mathrm{tr}(\tilde{\rho}) = \sum_{h=1}^{H} w_h \, \|u_h\|_2^2 \tag{52}$$

is used to define the normalized state

$$\rho = \frac{\tilde{\rho}}{T}, \tag{53}$$

which satisfies $\mathrm{tr}(\rho) = 1$ whenever $T > 0$.

The purity of ρ is given by

$$\Pi(\rho) = \mathrm{tr}(\rho^2) = \sum_k \lambda_k^2, \tag{54}$$

where $\{\lambda_k\}$ are the eigenvalues of ρ. Purity takes values in $(0, 1]$: it attains $\Pi(\rho) = 1$ if and only if ρ is a pure (rank-one) state, and decreases as ρ becomes more mixed. Substituting Eq. (53) into Eq. (54) and using the explicit form of $\tilde{\rho}$ yields

$$\Pi(\rho) = \frac{1}{T^2}\mathrm{tr}(\tilde{\rho}^2) = \frac{1}{T^2} \sum_{h=1}^{H} \sum_{h'=1}^{H} w_h w_{h'} \left|\langle u_h, u_{h'}\rangle\right|^2. \tag{55}$$

Thus, purity increases when a single head dominates and the active head states are strongly aligned, and decreases when many heads with diverse directions contribute, yielding a more mixed representation.

In the implementation we compute Eq. (52) and Eq. (55) in a fully vectorized form. Let

$$\gamma_{hh'} = \left|\langle u_h, u_{h'}\rangle\right|^2, \tag{56}$$

so that

$$T = \sum_h w_h \|u_h\|_2^2, \qquad \mathrm{tr}(\tilde{\rho}^2) = \sum_{h,h'} w_h w_{h'} \gamma_{hh'}, \tag{57}$$

and

$$\Pi(\rho) = \frac{\sum_{h,h'} w_h w_{h'} \gamma_{hh'}}{\left(\sum_h w_h \|u_h\|_2^2\right)^2}. \tag{58}$$

We store T as `StateTrace` and $\Pi(\rho)$ as `StatePurity`. Following standard practice, we also report an effective rank

$$R_{\mathrm{eff}}(\rho) = \frac{1}{\Pi(\rho)}, \tag{59}$$

which behaves like the effective number of active eigencomponents of ρ and is implemented as `StateEffRank`. As a purely classical baseline on the mixture weights, we additionally track

$$N_{\mathrm{eff}}(w) = \frac{1}{\sum_h w_h^2}, \tag{60}$$

the effective number of active heads, stored as `Neff`.

We use purity (and its derived effective rank) as a representation-level uncertainty indicator: high purity (low effective rank) corresponds to confident, sharply concentrated internal states, whereas low purity (high effective rank) signals that multiple interference heads with diverse responses are simultaneously active, indicating a more uncertain representation even when the output class probabilities may appear confident.

E.3 Coverage-based uncertainty from the reject channel

Finally, the POVM head provides an uncertainty signal at the level of *measurement coverage*. Recall that we implement a $(C+1)$-outcome POVM $\{E_1, \ldots, E_C, E_R\}$, where E_R is a dedicated reject effect. For a given mixed state ρ, the unnormalized class scores s_j and the reject score s_R can be written as

$$s_j = \mathrm{tr}(E_j \rho), \qquad s_R = \mathrm{tr}(E_R \rho), \tag{61}$$

and their sum

$$\mathrm{TraceScore} = \sum_{j=1}^{C} s_j + s_R \tag{62}$$

approximates $\mathrm{tr}(\rho)$ up to numerical regularization.

We define the *coverage* of the predictive sub-POVM as

$$\kappa = \frac{\sum_{j=1}^{C} s_j}{\sum_{j=1}^{C} s_j + s_R} \in (0, 1], \tag{63}$$

which measures the fraction of total POVM energy captured by the C task classes. Low coverage (small κ) indicates that a significant portion of the state is aligned with the reject direction, and thus that the sample lies outside the region where the task-specific effects provide a good description. In parallel, the normalized reject probability

$$p_R = \frac{s_R}{\sum_{j=1}^{C} s_j + s_R} \tag{64}$$

offers a more conventional reject option at the output level.

In the implementation, s_j and s_R are obtained from the raw scores `ScoreFull`, coverage κ is stored as `CoverageKappa`, and p_R is stored as `RejectProb`. These quantities are combined with the entropy- and purity-based indicators to form composite reject scores used in our experiments; however, all three uncertainty indicators (entropy/margin, purity/effective rank, and coverage/reject probability) are fully endogenous to the POVM head and can be inspected independently for analysis.

F Details of Modules

To facilitate a more detailed understanding of the design of the IEF module, we provide the framework diagrams for both components as illustrated in the Fig. A5.

G Additional Visualization and Case Studies

G.1 Modality Main Effects

The modality main effect characterizes the average dominant contribution of each modality to the Born-type score within a given class channel. Figure 3 visualizes this quantity under the four decision cases: each subfigure corresponds to one quadrant of the confusion matrix; rows represent the Real / Spoof class channels and their mean, columns correspond to the three modalities RGB / Depth / IR, and the values denote the normalized proportion of the modality main effects under the corresponding row condition.

We observe that in TA and FA (Fig. 3(a),(c)), i.e., cases where the final prediction is "Real", the main effects in the Real channel predominantly come from Depth and IR (together accounting for about

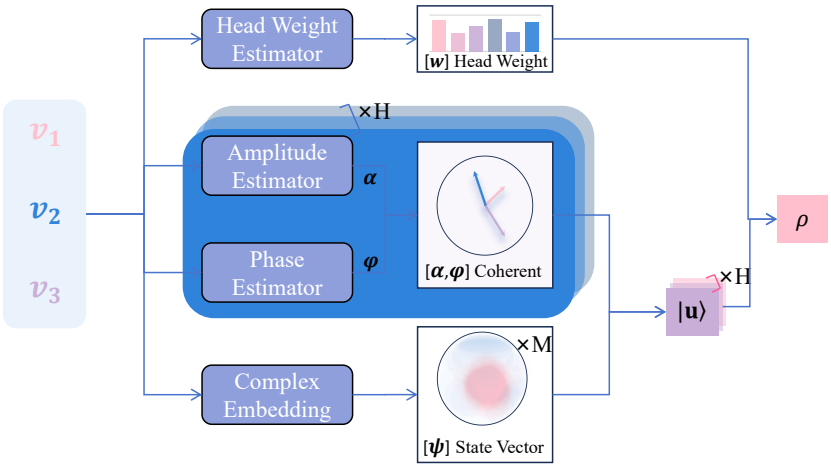

Fig. A5: Architecture of the Interference-based Evidence Fusion (IEF) module, showing the amplitude encoding, phase encoding, coherent superposition, and mixed-state construction stages.

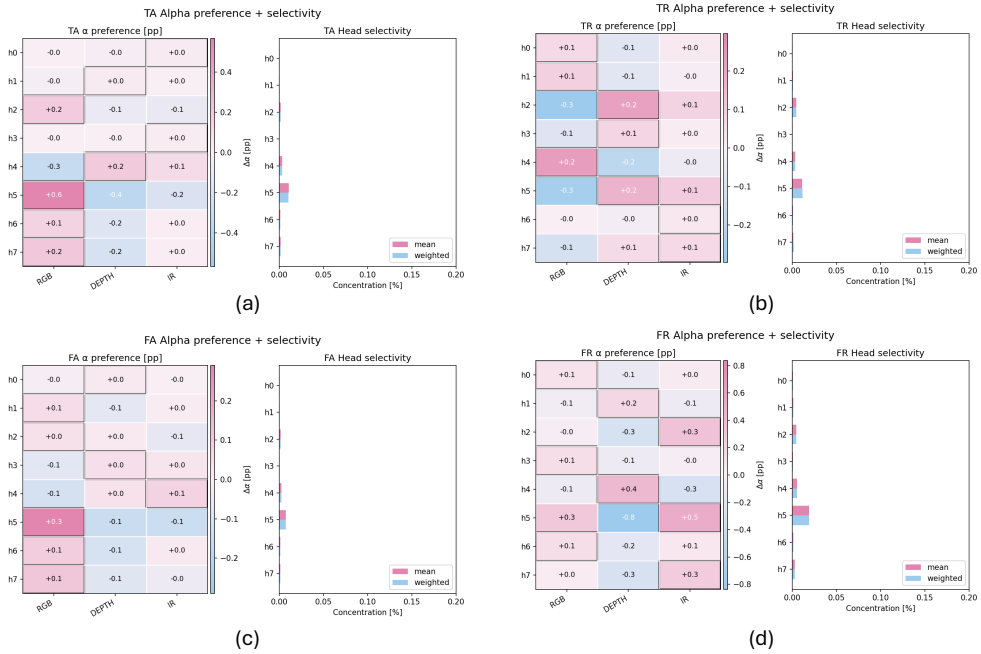

Fig. A6: Multi-head modality preferences: per-head deviations of amplitude weights over RGB/Depth/IR (heatmap) and the corresponding head selectivity scores (bar chart).

75%–80%), while RGB only contributes 18.2%/25.3%. In contrast, the residual response in the Spoof channel is highly concentrated in RGB (74.5%/61.2%), with substantially lower contributions from Depth and IR. This indicates that when the model decides to accept a face as real, it tends to rely on the geometric and spectral modalities to accumulate Real evidence, whereas attack-related texture artifacts are mainly captured in RGB and reflected in the Spoof-channel score.

In TR and FR (Fig. 3(b),(d)), i.e., cases where the final prediction is "Spoof", the division of roles among modalities exhibits a different structured pattern. In TR (correctly rejecting attacks), the main effects in the Spoof channel come more from Depth and IR (39.1% and 34.7%), with RGB contributing only 26.3%; conversely, the latent response in the Real channel is mainly concentrated in RGB (56.4%, with Depth/IR at 17.7%/26.0%), suggesting that in this regime, abnormal patterns

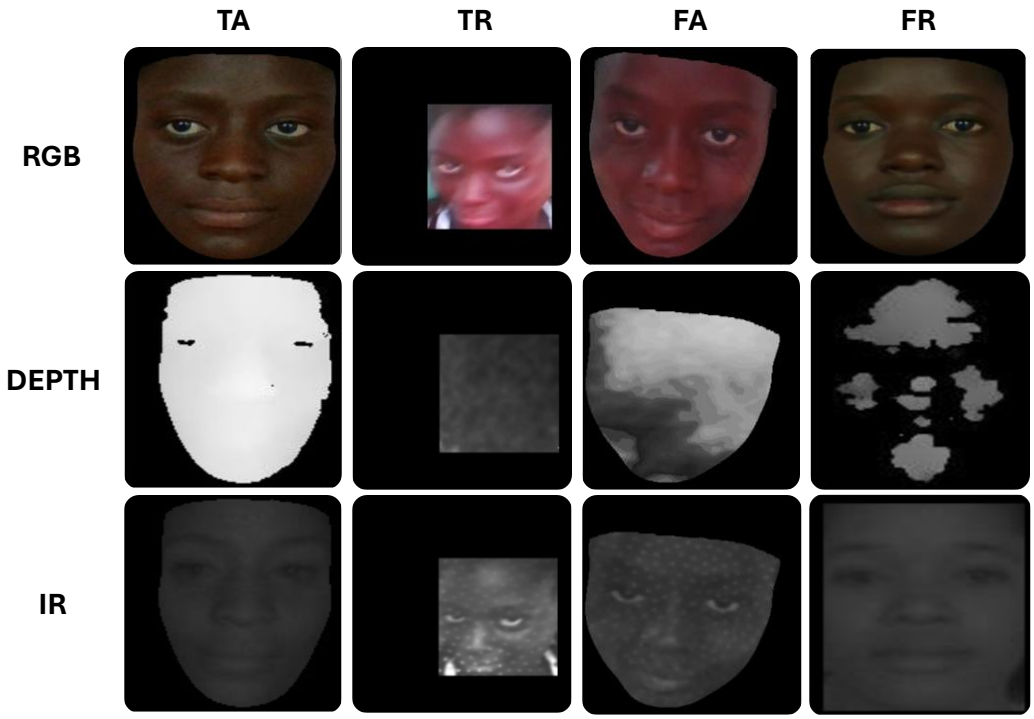

Fig. A7: Representative multimodal examples (RGB/Depth/IR) for the four decision types (TA/TR/FA/FR), illustrating typical correct and failure patterns of SEM-FAS. **Note:** Face images are from the CeFA (22) dataset.

in depth and infrared are key cues for distinguishing attacks, while RGB is more prone to drive a "looks like real" response. In FR (falsely rejecting real faces), the Real channel is almost entirely dominated by Depth (69.9%, with RGB and IR contributing only 21.4%/8.7%), whereas the Spoof channel exhibits high main effects on both IR and RGB (48.9% and 42.6%, with Depth at only 8.5%). In other words, falsely rejected real faces tend to show stronger attack-like characteristics in RGB/IR, or suffer from poor depth quality, such that the Spoof score is excessively amplified along these two modalities.

Overall, the main-effect proportions of all three modalities remain clearly non-zero across the four cases, indicating that the model does not collapse to a single-modality decision. Instead, it forms a collaborative pattern in which "one relatively dominant modality plus one or two auxiliary modalities" work together under different decision scenarios. These patterns are also broadly consistent with the physical characteristics of RGB, Depth, and IR and their expected behavior on real vs. attack samples, providing empirical support for using modality main effects as an explanatory signal of the decision process. Combined with the missing-modality results in Protocol 2, this further clarifies a more formal conclusion: across configurations, the performance degradation caused by removing the IR modality is systematically larger than that caused by removing Depth. In the full-modality setting, the Real channel relies heavily on the joint contribution of Depth and IR; once IR is removed, this combination is broken and Depth alone is insufficient to sustain the original discriminative capacity.

G.2 Modality Pairwise Effects

The modality pairwise effect captures whether different modalities act cooperatively or antagonistically within the same class score. As defined earlier, the values in Figure 4 correspond to a normalized and compressed "synergy" coefficient derived from the pairwise effects, ranging in $[-2, 2]$, where values near zero indicate approximate independence, positive values indicate constructive synergy, and negative values indicate predominantly suppressive interactions.

Across all four quadrants, a direct observation is that the synergy coefficients for all three modality pairs are almost always positive, with mean values concentrated around 0.6. This suggests that under the current training configuration, multimodal information is not being used by the model as "mutual veto" signals; instead, the modalities exhibit generally mild to moderate cooperative relationships. Among them, Depth IR tends to show slightly higher synergy than the two modality pairs involving RGB, which is consistent with the complementary nature of their geometric and spectral cues for real faces.

Focusing on each quadrant, we can see meaningful differences in synergy strength between the Real and Spoof channels. In the TA and FA plots, i.e., when the model ultimately predicts Real, the synergy coefficients of the three modality pairs are systematically higher in the Real channel than in the Spoof channel (e.g., in TA, Depth IR synergy is 0.65 in Real vs. 0.47 in Spoof), indicating that when the model tends to accept a face as real, the modalities are more inclined to provide mutually reinforcing evidence in the Real channel, while the synergy in the Spoof channel is weaker. In contrast, in the TR and FR plots, i.e., when the model ultimately predicts Spoof, the pattern is reversed: synergy in the Spoof channel is clearly stronger than in the Real channel (e.g., in TR, RGB IR synergy is 0.53 in Spoof vs. 0.27 in Real; in FR, RGB IR synergy is 0.58 in Spoof vs. 0.19 in Real), suggesting that when the model decides a face is an attack, multimodal evidence is more likely to be "amplified in the same direction" within the Spoof channel, while responses in the Real channel remain more dispersed.

Overall, these statistics show, on the one hand, that the quantum interference module is indeed learning signed interactions between modalities rather than performing a simple per-modality weighted sum; on the other hand, the consistent pattern of "stronger synergy in the predicted class channel and weaker synergy in the opposite class channel" across all four quadrants provides additional evidence for understanding the model's multimodal fusion behavior from an interaction-based perspective. Quantitatively, across all four decision types the normalized synergy coefficients of all three modality pairs remain consistently positive and are concentrated around a moderate magnitude (cf. Fig. 4), confirming that the SEM-FAS fusion module systematically learns cooperative cross-modal interactions rather than relying on a trivial per-modality weighted sum.

G.3 Multi Head Modality Preferences

Figure A6 analyzes "which head prefers which modality" inside the multi head interference module. The heatmap on the left shows, for each head, its preference over different modalities, measured in percentage points (pp): each cell can be interpreted as the deviation of that head's average amplitude weight on a given modality from the global mean level. Positive values indicate a relative preference for that modality, while negative values indicate a relative suppression. The small bar chart on the right shows the head selectivity, i.e., the modality concentration of each head; larger values mean the head's weight distribution is more "biased" toward a particular modality, while smaller values indicate a more uniform allocation across the three modalities.

Overall, most cells across the four quadrants are very close to zero, and the head selectivity scores fall into a relatively low range. This suggests that under the current training configuration, the interference heads are generally still "multimodal": they mainly make mild adjustments to the allocation over RGB, Depth, and IR, rather than forming extreme "single modality heads." In the TA and FA settings, i.e., cases where the final decision is Real, we can see that head h_5 shows a noticeable positive preference for RGB (about +0.6 pp in TA and +0.3 pp in FA), with corresponding negative deviations on Depth and IR; the other heads exhibit weaker and sign alternating preferences. This is consistent with the earlier observation from the modality main effects that "attack cues are primarily concentrated in RGB," and it suggests that h_5 is more sensitive to RGB direction changes in these two quadrants, while the remaining heads perform a more balanced fusion over the three modalities.

In the TR and FR settings, i.e., when the final decision is Spoof, the modality preference pattern becomes slightly different. In TR, several heads show small positive deviations on Depth and IR and slightly negative deviations on RGB; in FR, head h_4 exhibits a clearer preference for Depth, and head h_5 shows simultaneous positive deviations on both RGB and IR, indicating that in these cases some heads lean mildly toward geometric or infrared cues. The head selectivity curves on the right show that even in these situations, the concentration levels are still far from a regime where "a head almost exclusively relies on a single modality," indicating that these are mild tendencies built on a multimodal foundation rather than strict separations.

Putting together the modality main effects, pairwise synergy, and this multi head preference visualization, we see that the interference heads in SEM-FAS do not assign weights arbitrarily; instead, they form a set of coordinated modality subpreferences: a few heads place slightly more emphasis on certain modalities in specific decision quadrants, in line with the trends observed from the Born score based modality main effects, while the remaining heads maintain a more balanced multimodal fusion. This structure avoids collapsing decisions onto a single modality, while allowing certain heads to carry "extra sensitivity" on particular modalities, thereby introducing fine grained modality bias without sacrificing overall robustness. Combined with the ablation results on quantum-inspired state modeling (where mixed states overall outperform single head pure states) and the medium strength synergies observed in the pairwise effects, this supports the view that multi head classical mixing is indeed used by the model to learn structured modality interactions, rather than merely implementing a simple per modality weighting, which in turn helps explain why the mixed state configuration outperforms the pure state configuration in cross domain testing.

The small but systematic deviations in the heatmaps, together with the low to moderate selectivity scores in Fig. A6, therefore provide a quantitative characterization of how different heads implement mild, structured modality preferences on top of a predominantly multimodal representation.

G.4 Case Study

Figure A7 presents a representative example of the four decision types TA/TR/FA/FR, showing their multimodal appearances, to be compared with the earlier visualizations of modality main effects and pairwise effects. Vertically, the rows correspond to RGB, Depth, and IR; horizontally, the columns from left to right correspond to TA, TR, FA, and FR.

In the TA column, all three modalities appear relatively clean: skin tone and texture look natural in RGB, Depth provides a complete and smooth facial contour, and facial structures are clearly visible in IR. Combined with the preceding statistics, this is consistent with the observation that, in TA cases, the modality main effects in the Real channel are primarily supported by Depth and IR, and all modality pairs exhibit positive synergy. The configuration of "normal texture with reliable geometric and infrared cues" in this example aligns well with the model's strong reliance on Depth/IR in the Real channel, making it easy for the sample to be correctly accepted.

The TR column shows an attack sample that is correctly rejected. The Depth modality does not show a clear facial shape, instead exhibiting low quality planar depth noise; the IR image also contains patterns inconsistent with a real face. Such cross-modal, jointly abnormal patterns match the earlier finding that, in TR cases, the Spoof channel's main effects are dominated by Depth and IR. This indicates that, in this quadrant, the model relies more on abnormal geometric and infrared cues to accumulate Spoof evidence and therefore produces a stable reject decision for this type of attack.

The FA column corresponds to an attack sample that is incorrectly accepted. Here, the facial texture appears relatively natural in RGB, while the IR image shows dense speckle-like patterns, and the overall appearance is closer to a print attack. From the modality main effect and pairwise synergy visualizations, we observed that in FA cases the Real channel also mainly relies on Depth and IR, with generally positive synergy across modality pairs. This means that when an attack looks sufficiently "realistic" along the geometric and infrared dimensions, multimodal synergy in the Real channel can amplify these cues and overshadow residual spoof artifacts in RGB, eventually causing a false accept. This example illustrates that attacks which are cross-modally consistent and globally biased toward a "real" appearance are particularly prone to induce FA.

The FR column shows a real sample that is incorrectly rejected. The RGB image appears mostly normal, but the Depth modality exhibits fragmented, sparse patches, and the IR image has a blurry facial outline with very low contrast. The three modalities thus disagree noticeably on whether a stable facial geometry and spectral structure is present. Combined with the earlier analysis of main effects and synergy in FR cases, this suggests that such real samples tend to resemble anomalous patterns in the Depth/IR channels, thereby accumulating strong joint evidence in the Spoof channel while weakening the Real channel, leading to a false reject.

Overall, this set of cases ties the visually observable multimodal quality to the modality main effects and pairwise synergy extracted from SEM-FAS, showing that these explanatory signals are broadly consistent with human intuition about sample difficulty and cross-modal consistency. Together with the Protocol 2 findings, that removing IR leads to systematically larger performance degradation than

removing Depth, and that in some cross-dataset combinations single-modality configurations can even outperform multimodal ones, these examples highlight two points: on the one hand, Depth-IR indeed provides crucial joint geometric and spectral evidence for the Real channel under full-modality conditions; on the other hand, when multiple modalities in the target domain simultaneously drift in the wrong direction, pairwise synergy in the predicted class channel can amplify these shifts, thereby weakening the advantage of multimodal over single-modal methods in certain cross-domain settings and offering a concrete, visual mechanism for the phenomena observed in Protocol 2.

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: [Yes]

Justification: The paper includes theoretical results, and the full assumptions and proofs are provided in the appendix.

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.