CALIBRATE ELSEWHERE, VERIFY HERE: DISJOINT IN-SCENE CALIBRATION FOR MULTIMODAL FACE ANTI-SPOOFING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal face presentation attack detection must distinguish presentation effects from ordinary changes in illumination and sensing. Target observations can contain acquisition cues, but when acquisition and attack effects overlap, target-only calibration is ambiguous: marginalization preserves uncertainty without independent sample-specific evidence, while point adaptation can absorb nuisance-aligned attack evidence. We introduce SCENECAL, which learns type-conditioned responses of heterogeneous non-target regions to a shared acquisition state and conditions the face test on control-derived evidence. We analyze plug-in attenuation, target-only marginalization, and disjoint control conditioning in one Gaussian model. On public WMCA RGB–Depth leave-one-attack-out, SCENECAL reaches 2.99% ACER, versus 3.37% for CTNet and 3.54% for MMDG++; on HQ-WMCA it reaches 9.21%, versus 9.88% and 10.23%. On HQ-MMFAS, CTNet is stronger on known instruments (2.06% vs. 3.08%), while SCENECAL reaches 4.72% on held-out instruments, compared with 5.08% for CTNet and 5.45% for DADM.

1 INTRODUCTION

Multimodal face presentation attack detection (PAD) combines physical measurements such as RGB appearance, scene depth, and near-infrared (NIR) response (Yu et al., 2023b; Zhang et al., 2019; Liu et al., 2021). A replay display can reproduce visible appearance while remaining a planar emissive surface, whereas a silicone mask can preserve facial shape while responding differently from living skin under infrared illumination. These differences make multiple sensing channels valuable for attacks that remain plausible in RGB alone (George et al., 2019; Heusch et al., 2020).

The expected cross-modal relation is nevertheless capture dependent. Illumination, exposure, sensing range, active-IR strength, and depth quality vary from sample to sample. Recent multimodal methods address unreliable sensing and modality-dependent domain shifts through adaptive fusion, balancing, and alignment (Liu et al., 2023; Lin et al., 2024; Yang et al., 2025; Chong et al., 2026). A verifier must distinguish acquisition variation, which is shared by the scene, from a target-specific change caused by the presentation medium. With a calibrated reference in every frame this distinction would be straightforward, but ordinary PAD scenes contain no material of known response.

The face can itself contain cues about illumination or sensing state, so target-only estimation is not intrinsically impossible. The difficulty is that the same observations must separate acquisition variation from presentation changes when the two occupy overlapping response directions. Marginalization accounts for state uncertainty but adds no independent evidence about the current capture; a point estimate fitted to the target can reduce uncertainty yet absorb part of an attack. Related instance-aware and test-time methods likewise estimate sample variation from the target (Zhou et al., 2023; 2024). External controls are useful because they provide additional sample-specific evidence without requiring the target to resolve this ambiguity by itself.

The surrounding scene provides such information whenever some non-target surfaces respond to the same acquisition factors. The useful signal is not homogeneity. Hair, fabric, and background surfaces have different baseline responses, and these differences are informative because they probe the shared capture through different response directions. For example, active-IR response varies jointly with surface reflectance and range; observing several materials at different distances can constrain the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sensing state more strongly than repeatedly observing one material. In our experiments, eight controls spanning at least four learned types reduce posterior trace by 17% relative to sixteen controls drawn from one dominant type.

We introduce SCENECAL, a bona-fide-trained model that learns type-conditioned responses of heterogeneous scene regions to a shared low-dimensional acquisition state. Non-target controls outside an expanded face guard define a sample-specific state posterior, and the face is evaluated under the corresponding predictive covariance. When controls are scarce, redundant, or partially observed, predictive uncertainty adapts to the evidence available to the verifier.

Our analysis distinguishes plug-in adaptation, target-only marginalization, and disjoint control conditioning, showing how each treats nuisance-aligned attack evidence and acquisition uncertainty. We make three contributions: (i) heterogeneous in-scene controls with learned type-conditioned responses; (ii) a unified probabilistic analysis of target-only and disjoint calibration; and (iii) mechanism tests, recent multimodal PAD comparisons, held-out-instrument evaluation, transfer, and sensor-shift experiments.

2 RELATED WORK

2.1 MULTIMODAL FACE ANTI-SPOOFING

Multimodal PAD combines complementary sensing cues from RGB, depth, NIR, thermal, or SWIR observations (Zhang et al., 2019; Liu et al., 2021; George et al., 2019; Mostaani et al., 2020; Heusch et al., 2020). Recent models use flexible transformers, reliability-aware balancing, transfer learning, and domain–modality alignment to improve robustness when sensing quality or domain statistics vary (Liu et al., 2023; Yu et al., 2023a; Lin et al., 2024; 2025; Yang et al., 2025; Zhang & Li, 2025; Chong et al., 2026; Zhang & Li, 2026). Domain-generalization and test-time methods further address unseen environments through separability, gradient alignment, instance-aware adaptation, and test-time projection (Sun et al., 2023; Le & Woo, 2024; Zhou et al., 2023; 2024). These methods adapt representations or fusion rules. SCENECAL instead estimates the acquisition state of the current capture from non-target content and evaluates the face only after that state is fixed.

2.2 LOCAL EVIDENCE AND SOURCE CONSISTENCY

Local PAD and forensic methods exploit patch-level presentation cues, local-to-global facial structure, or source consistency (Wang et al., 2022; Hong et al., 2024; Huh et al., 2018; Mayer & Stamm, 2020; Zhao et al., 2021; Wang et al., 2025). Such methods establish that authenticity evidence can be spatially non-uniform, but source consistency alone does not specify how a bona-fide material should respond under the current sensing condition. Our controls are therefore not treated as additional attack evidence. They are used to infer a shared acquisition variable that conditions the target test.

2.3 LEARNING FROM CONTROLS

Negative controls, RUV, and surrogate-variable analysis estimate unwanted variation from variables that do not carry the target effect (Lipsitch et al., 2010; Gagnon-Bartsch & Speed, 2012; Leek & Storey, 2007). RX detection uses surrounding background statistics, while color constancy and flashing-based liveness exploit illumination or reflectance references (Reed & Yu, 1990; Gijsenij et al., 2011; Tang et al., 2018). Conditional neural processes infer target distributions from contextual observations (Garnelo et al., 2018). Our setting differs in that the controls are heterogeneous and their responses are neither known nor homogeneous: the type-conditioned response itself must be learned from bona-fide scenes.

3 METHOD

3.1 SETUP

Figure 1 provides an overview of our approach. We consider an aligned multimodal capture with RGB as the reference stream and one or more additional sensing streams. Public WMCA and HQ-

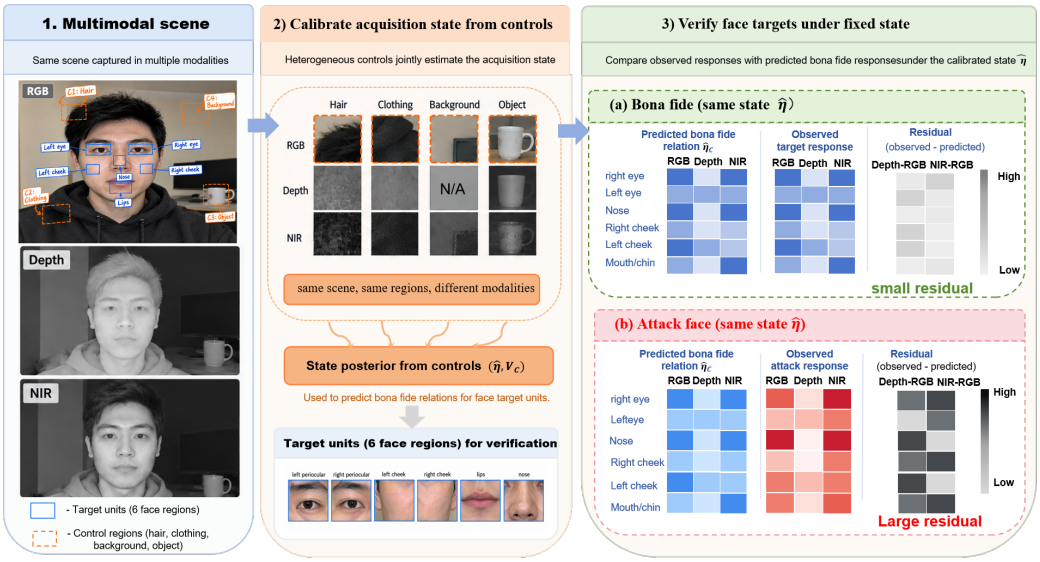

Figure 1: **SceneCal overview.** Non-target RGB–Depth–NIR regions calibrate a shared acquisition state. With this state fixed, the face is tested against its model-implied bona-fide response. Schematic illustration.

WMCA experiments use RGB–Depth to match the published comparison protocols; HQ-MMFAS uses RGB–Depth–NIR. Each frame contains six fixed target units—left/right cheek, nose, mouth, and left/right periocular regions—defined from landmarks and parsing masks. Candidate controls are superpixels outside a $1.6\times$ expanded face guard.

The formulation assumes that some non-target content remains outside the presentation support, shares acquisition factors with the face, and is not systematically altered by the target presentation mechanism. This is an exclusion-style assumption on the physical effect of the attack, not statistical independence from the attack label; local violations motivate the robust inference in Section 3.4.

For unit i, c_i denotes a low-capacity type code and p_i a geometry descriptor. For controls, p_i contains normalized image position, log-area, median valid depth, and a depth-validity indicator. For targets, measured depth is never used to condition the reference model; p_i contains only landmark-derived position and scale, with the depth fields masked out.

Let $y_i \in \mathbb{R}^{d_y}$ be the cross-modal relation for unit i. Bona-fide data follow

$$y_i \mid \eta_x \sim \mathcal{N}(\mu_i + B_i \eta_x, \Sigma_{c_i}), \qquad \eta_x \sim \mathcal{N}(0, \mathbf{I}_q), \tag{1}$$

where $\mu_i = \mu_\theta(c_i, p_i)$ and $B_i = B_\theta(c_i, p_i) \in \mathbb{R}^{d_y \times q}$. The diagonal $\Sigma_{c_i} = \mathrm{diag}(\exp s_\theta(c_i))$ models type-dependent residual variability. The isotropic prior fixes latent scale; a common orthogonal rotation of η_x and the loading family leaves all predictive quantities unchanged.

3.2 Cross-modal relation

The relation representation removes three nuisance freedoms before the latent model is learned. First, for each modality and type, we subtract the source-training type mean and fit PCA to the pooled *within-type* residuals; this prevents the low-dimensional budget from being dominated by face-versus-background semantics. Second, the retained coordinates are whitened, fixing a common metric within each modality. After whitening, the remaining linear basis ambiguity is orthogonal, so each non-RGB modality is aligned to RGB with an orthogonal Procrustes map Q_m estimated from paired bona-fide units.

Let $\tilde{z}_i^{(m)} \in \mathbb{R}^{d_f}$ denote the resulting type-centered, whitened feature for modality m. For modality set $\mathcal{M}$ with RGB as reference,

$$y_i = \mathrm{concat}_{m \in \mathcal{M} \setminus \{\mathrm{RGB}\}} \left(Q_m \tilde{z}_i^{(m)} - \tilde{z}_i^{(\mathrm{RGB})} \right). \tag{2}$$

We use $d_f = 32$, giving $d_y = 32$ for RGB–Depth and $d_y = 64$ for RGB–Depth–NIR. The transformation is fixed from the source bona-fide split. Because type centering uses source-training means, a cross-domain type-mean shift enters the control-model bias analyzed in Section 4. CCA gives similar performance within this relation family. Direct concatenation contains at least as much raw information, but it is a weaker inductive bias for the rank-q acquisition model because high-variance identity and pose directions dominate the fit; Section 5.5 quantifies both effects.

Target types are fixed semantic categories. Control types are obtained by assigning a frozen RGB descriptor to one of $K = 32$ source-domain prototypes; only the discrete code and geometry are passed to μ_θ and B_θ. Photometric augmentation is used when fitting the codebook so that the type code remains stable under the acquisition changes represented by η_x.

3.3 Partial observations and disjoint calibration

Depth and NIR may be invalid on part of a region. A binary selection matrix O_i keeps the observed relation dimensions, giving $\bar{y}_i = O_i y_i$, $\bar{\mu}_i = O_i \mu_i$, $\bar{B}_i = O_i B_i$, and $\bar{\Sigma}_i = O_i \Sigma_{c_i} O_i^\top$. Controls with no observed relation dimensions are omitted. In RGB–Depth, a control with no valid depth relation contributes no information; in RGB–Depth–NIR, a unit can still contribute through its valid relation block.

For a control set $\mathcal{C}$, define

$$J_{\mathcal{C}} = \sum_{i \in \mathcal{C}} \bar{B}_i^\top \bar{\Sigma}_i^{-1} \bar{B}_i, \qquad V_{\mathcal{C}} = (\mathbf{I}_q + J_{\mathcal{C}})^{-1}, \tag{3}$$

with posterior mean

$$\hat{\eta}_{\mathcal{C}} = V_{\mathcal{C}} \sum_{i \in \mathcal{C}} \bar{B}_i^\top \bar{\Sigma}_i^{-1} (\bar{y}_i - \bar{\mu}_i). \tag{4}$$

The face does not enter either expression.

Stacking the observed dimensions of the six target units gives $\bar{y}_\mathcal{T}$, $\bar{\mu}_\mathcal{T}$, and $\bar{B}_\mathcal{T}$ with block-diagonal observation covariance $R_\mathcal{T}$. The posterior predictive distribution is

$$\bar{y}_\mathcal{T} \mid \bar{y}_\mathcal{C} \sim \mathcal{N}\big(\bar{\mu}_\mathcal{T} + \bar{B}_\mathcal{T} \hat{\eta}_\mathcal{C}, \Omega_{\mathcal{T}|\mathcal{C}}\big), \quad \Omega_{\mathcal{T}|\mathcal{C}} = R_\mathcal{T} + \bar{B}_\mathcal{T} V_\mathcal{C} \bar{B}_\mathcal{T}^\top. \tag{5}$$

Let $d_\mathcal{T}$ be the number of observed scalar target dimensions after applying the target availability masks; its nominal value is 192 for RGB–Depth and 384 for RGB–Depth–NIR. We first compute the predictive quadratic

$$Q(x) = r_\mathcal{T}^\top \Omega_{\mathcal{T}|\mathcal{C}}^{-1} r_\mathcal{T}, \qquad r_\mathcal{T} = \bar{y}_\mathcal{T} - \bar{\mu}_\mathcal{T} - \bar{B}_\mathcal{T} \hat{\eta}_\mathcal{C}. \tag{6}$$

Under the clean Gaussian model, $Q(x) \sim \chi^2_{d_\mathcal{T}}$ for bona-fide samples, including when $d_\mathcal{T}$ varies because of missing target dimensions. For the clean Gaussian model, this gives the null-calibrated score

$$S(x) = -\log\Big[\bar{F}_{\chi^2_{d_\mathcal{T}}}(Q(x))\Big],$$

where $\bar{F}$ is the chi-square survival function; under the reference model, S follows an $\mathrm{Exp}(1)$ distribution independently of $d_\mathcal{T}$. The Gaussian log-determinant is used in the training likelihood but not in the detection score. The predictive quadratic is evaluated with a $q \times q$ Woodbury solve (Appendix A.3). Video scores average 16 uniformly sampled frame scores; the $\mathrm{Exp}(1)$ reference applies to individual frames, while video thresholds are selected on the development split.

For mechanism comparisons we use four related estimators. *Population* fixes $\eta = 0$ and uses only the type-conditioned mean and observation covariance. *Self-plugin* fits a point state from the target and scores the remaining residual as if that state were known. *Marginal* integrates the acquisition prior without using scene controls. *Joint-plugin* fits one point state from target and controls. SceneCal conditions only on controls and propagates their posterior uncertainty through Eq. (5).

3.4 Robust candidate-control inference

Candidate controls can be contaminated by presentation hardware, sensor failure, or a source-to-target material shift. We therefore use three Huber iteratively reweighted updates. At iteration t, let d_i

be the number of observed relation dimensions for control i and define the dimension-normalized residual

$$e_i^{(t)} = \left[\frac{(\bar{y}_i - \bar{\mu}_i - \bar{B}_i \hat{\eta}^{(t)})^\top \bar{\Sigma}_i^{-1} (\bar{y}_i - \bar{\mu}_i - \bar{B}_i \hat{\eta}^{(t)})}{d_i} \right]^{1/2}.$$

The Huber weight is $w_i^{(t)} = \min\{1, \tau / e_i^{(t)}\}$, so controls with different availability patterns are compared on a per-observed-dimension scale. The robust information matrix and state estimate replace each control contribution in Eqs. (3)–(4) by $w_i^{(t)}$ times that contribution. Because the weights depend on the observations, the resulting $V_{\mathcal{C}}^{(w)}$ is a robust uncertainty proxy rather than an exact Gaussian posterior covariance; downweighting reduces information and therefore increases this proxy in the intended direction. The chi-square calibration of Eq. (6) is exact for the clean Gaussian model and is used as a reference calibration after robust reweighting; its empirical bona-fide calibration is reported inAppendix G.

Training uses the same three-step robust inference. In 30% of training episodes, 10% of inference controls are replaced by mismatched bona-fide controls from another capture or perturbed with large empirical residuals. The Huber threshold is fixed from the 95th percentile of source-training dimension-normalized Mahalanobis control residuals.

3.5 Bona-fide episodic learning

The response model is learned without presentation attacks. For each bona-fide sample, controls are split into inference controls $\mathcal{C}_{\text{in}}$ and held-out controls $\mathcal{C}_{\text{held}}$. The state estimate from $\mathcal{C}_{\text{in}}$ predicts both the target and held-out controls. Training minimizes their joint predictive negative log-likelihood,

$$\mathcal{L}_{\text{pred}} = -\log p_\theta(\bar{y}_{\mathcal{T}}, \bar{y}_{\mathcal{C}_{\text{held}}} \mid \bar{y}_{\mathcal{C}_{\text{in}}}), \tag{7}$$

with control count and type composition varied across episodes. The loading functions are initialized from a rank-q factor model of bona-fide relation residuals and then optimized end-to-end with Eq. (7).

4 Analysis

We analyze the clean Gaussian model; robust reweighting is evaluated experimentally. Let R and B denote the stacked target observation covariance and loading matrix. Whiten the target with $R^{-1/2}$, define $C = R^{-1/2}B$ and $\tilde{\delta} = R^{-1/2}\delta$, and let $C = USV^\top$ have singular values s_j. Write $\tilde{\delta} = Ua + \delta_\perp$ with $\delta_\perp \perp \text{range}(U)$, and let $k = \text{rank}(C)$.

4.1 Target-only processing

A flat-prior plug-in fit removes the component of the target residual in $\text{range}(C)$. Its quadratic statistic has

$$Q_{\text{plugin}} \sim \chi^2_{d_{\mathcal{T}}-k}(\Lambda_{\text{plugin}}), \qquad \Lambda_{\text{plugin}} = \|\delta_\perp\|_2^2. \tag{8}$$

The aligned component Ua is absent from the score.

Marginalizing the latent acquisition state instead gives covariance $\mathbf{I} + CC^\top$ in whitened coordinates and

$$Q_{\text{marg}} \sim \chi^2_{d_{\mathcal{T}}}(\Lambda_{\text{marg}}), \qquad \Lambda_{\text{marg}} = \|\delta_\perp\|_2^2 + \sum_j \frac{a_j^2}{1+s_j^2}. \tag{9}$$

Proposition 1. Under the linear-Gaussian model, scoring the target-derived posterior-mean residual with its *in-sample residual covariance* is equivalent to target-only marginalization (Appendix A.1). A flat-prior plug-in residual score discards the aligned contribution $\sum_j a_j^2/(1+s_j^2)$ from marginal noncentrality. The plug-in statistic also has k fewer null degrees of freedom, so power comparisons use the exact null distribution of each statistic rather than noncentrality alone. In our setting $d_{\mathcal{T}} \gg q$; the mechanism experiment uses these estimator-specific degrees of freedom explicitly.

4.2 INDEPENDENT CONTROLS ADD INFORMATION

Controls give $V_{\mathcal{C}} = (\mathbf{I} + J_{\mathcal{C}})^{-1} \preceq \mathbf{I}$. The target statistic conditioned on clean controls has

$$Q_{\text{ctrl}} \sim \chi^2_{d_{\mathcal{T}}}(\Lambda_{\text{ctrl}}), \quad \Lambda_{\text{ctrl}} = \tilde{\delta}^\top (\mathbf{I} + C V_{\mathcal{C}} C^\top)^{-1} \tilde{\delta}. \tag{10}$$

Since $V_{\mathcal{C}} \preceq \mathbf{I}$, $\Lambda_{\text{ctrl}} \geq \Lambda_{\text{marg}}$. The two statistics have the same null degrees of freedom, so fixed-false-positive power is monotone in this increase.

Proposition 2. Conditioning on independent controls cannot reduce conditional detectability relative to target-only marginalization under the model. The inequality is strict when the controls contract a nuisance direction that affects the target and overlaps with the attack deviation. The per-direction form (Appendix A.3) shows that control gain depends on attack alignment a_j, nuisance-to-noise strength s_j^2, and posterior contraction. A nuisance direction that is weakly constrained by controls has posterior variance near its prior value and therefore reverts smoothly toward the Marginal case rather than failing the entire sample. We convert these noncentralities to analytic AUC using each estimator's own central/noncentral chi-square pair.

Control-model bias. If controls have a systematic source-to-target mean mismatch m_i, the inferred state is shifted by $b_\eta = V_{\mathcal{C}} \sum_i B_i^\top \Sigma_i^{-1} m_i$ in the fully observed notation. The target attack mean becomes $\delta - B_{\mathcal{T}} b_\eta$, giving

$$\Lambda_{\text{ctrl,bias}} = (\tilde{\delta} - C b_\eta)^\top (\mathbf{I} + C V_{\mathcal{C}} C^\top)^{-1} (\tilde{\delta} - C b_\eta).$$

This term motivates robust controls, condition-matched donor experiments, and reporting the state-norm shift under domain transfer.

The additive δ model is a local representation of an attack. If a presentation medium also changes the target loading, $B_{\mathcal{T}} \to B_{\mathcal{T}} + \Delta B$, then conditional on the current state its first-order deviation is $\delta(\eta) = \delta_0 + \Delta B \eta$, and the same analysis applies pointwise in η.

5 EXPERIMENTS

5.1 DATASETS, PROTOCOLS, AND EVALUATION

WMCA contains 1,941 recordings from 72 identities; we use its RGB–Depth seen and seven-attack leave-one-attack-out (LOO) protocols (George et al., 2019). HQ-WMCA contains 2,904 recordings, including 555 bona-fide recordings from 51 participants (Mostaani et al., 2020; Heusch et al., 2020). Following the public comparison setting, both datasets are evaluated with RGB–Depth. HQ-MMFAS contains 200 subjects, five lighting conditions, and thirteen physical presentation instruments captured with synchronized D435i RGB, depth, and NIR. Its train, development, and test partitions are subject- and session-disjoint. Protocol A evaluates known instrument families, whereas Protocol B additionally holds out instrument identities and mask donors.

Type means, PCA/whitening, Procrustes maps, control codebooks, and response models are fitted only on source-training bona-fide data in every fold; no test-fold information enters preprocessing. WMCA/HQ-WMCA use the development BPCER$= 1\%$ threshold and HQ-MMFAS uses the source-development EER threshold. Every matched comparison uses the same split, modalities, and threshold rule. Main performance tables focus on methods published from 2023 onward. We report ACER at these operating points, cross-dataset HTER, three-run variation, and paired uncertainty in Appendix I.

5.2 ACQUISITION STATE AND MECHANISM

On HQ-MMFAS, the control-derived state predicts lighting with 84.2% accuracy and exposure/NIR-gain severity with $R^2 = 0.78$. More importantly, a paired state-sensitivity test distinguishes acquisition variation from scene semantics. Changing lighting/exposure for the same subject gives a median normalized state distance of 1.18, whereas condition-matched bona-fide controls from different scenes/materials yield only 0.31 with cosine similarity 0.93; acquisition-mismatched controls yield 1.24 and 0.56, respectively. Identity probes on the control-derived state and target residual are only 0.8 and 1.6 points above chance. Additional state and calibration diagnostics are reported in Appendix F.

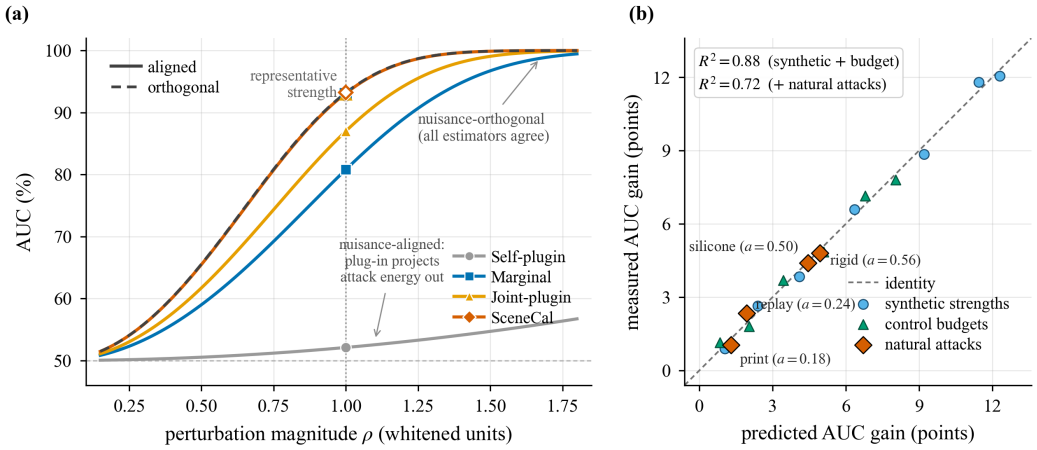

Figure 2: **Mechanism.** (a) AUC under aligned and orthogonal perturbations. (b) Analytic versus measured AUC gain.

Table 1: Recent RGB–Depth PAD comparisons, ACER (%) ↓. †: matched-protocol reimplementation with the same modalities, split, sampling, and threshold rule.

Method	WMCA seen	WMCA LOO	HQ-WMCA LOO
FM-ViT† (Liu et al., 2023)	1.42	4.18	11.42
Suppress & Rebalance† (2024) (Lin et al., 2024)	1.33	3.67	10.71
MMDG++† (Lin et al., 2025)	1.26	3.54	10.23
DADM† (Yang et al., 2025)	1.51	3.73	10.48
CTNet† (Chong et al., 2026)	**1.19**	3.37	9.88
CA-MMDG† (Zhang & Li, 2026)	1.28	3.46	10.05
SCENECAL	1.68	**2.99**	**9.21**

Figure 2 tests the analysis. Under nuisance-aligned perturbations, Self-plugin, Marginal, Joint-plugin, and SCENECAL obtain 52.1, 80.8, 87.0, and 93.1% AUC; under orthogonal perturbations all are near 93%. Analytic gain explains $R^2 = 0.88$ of measured gain across perturbation strengths and control budgets, and 0.72 after adding natural attacks. Print, replay, silicone, and rigid-mask alignment fractions are 0.18, 0.24, 0.50, and 0.56.

5.3 PUBLIC RGB–DEPTH BENCHMARKS

Table 1 compares recent methods under matched RGB–Depth protocols. SCENECAL obtains 2.99% WMCA LOO and 9.21% HQ-WMCA LOO ACER. Against CTNet, paired video bootstrap gives ΔACER $= -0.38$ points (95% CI $[-0.69, -0.09]$) on WMCA and -0.67 ($[-1.31, -0.12]$) on HQ-WMCA. Three-run seed SD is 0.17/0.34 ACER. The hardest WMCA folds remain flexible-mask and glasses (6.00/10.80 ACER); complete fold and operating-point results are in Appendices H and J.

5.4 CONTEXT EVIDENCE VERSUS CALIBRATION

Table 2 separates context from calibration under the same HQ-MMFAS RGB–Depth–NIR splits. A target+context supervised head improves Protocol B from 8.91 to 7.54% ACER, while SCENECAL reaches 4.72%. On known instruments, CTNet and MMDG++ remain stronger (2.06/2.21% vs. 3.08%), whereas the held-out setting favors disjoint calibration. Within our model family, Self-plugin changes from 4.68 to 10.91% across A/B, while Marginal, Joint-plugin, and SCENECAL progressively improve held-out performance.

Table 2: HQ-MMFAS RGB–Depth–NIR ACER (%) ↓. † denotes our matched-protocol reimplementation.

Method	Protocol A	Protocol B
Target-only supervised head	2.34	8.91
Target+context supervised head	2.28	7.54
FM-ViT† (Liu et al., 2023)	2.50	7.86
Suppress & Rebalance† (2024) (Lin et al., 2024)	2.29	6.63
MMDG++† (Lin et al., 2025)	2.21	5.66
DADM† (Yang et al., 2025)	2.49	5.45
CTNet† (Chong et al., 2026)	**2.06**	5.08
CA-MMDG† (Zhang & Li, 2026)	2.57	5.83
One-class SVM	5.92	10.84
Population	5.10	10.15
Self-plugin	4.68	10.91
Marginal	3.91	7.42
Joint-plugin	3.46	5.96
SCENECAL	3.08	**4.72**

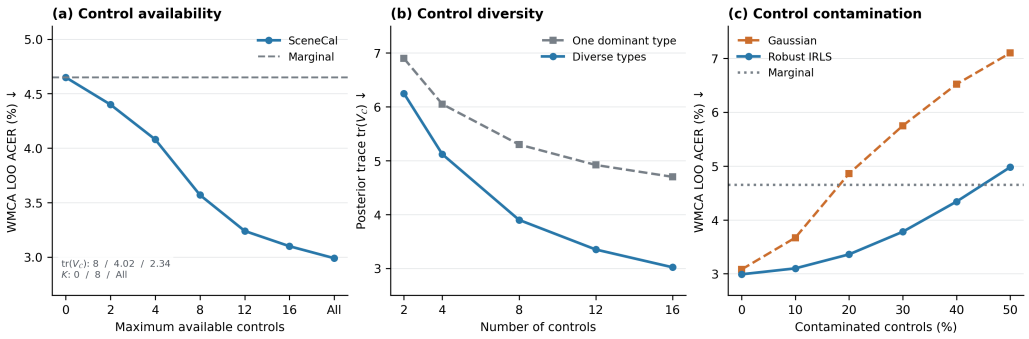

Figure 3: **Control assumptions and applicability.** (a) Fewer controls move SCENECAL toward Marginal. (b) Type diversity improves posterior contraction. (c) Robust IRLS delays degradation under increasing control contamination.

5.5 HETEROGENEOUS CONTROLS AND APPLICABILITY

WMCA, HQ-WMCA, and HQ-MMFAS provide on average 15.1, 22.4, and 25.8 usable controls per frame, with median information ranks 5.5, 6.4, and 7.1 of $q = 8$. Eight controls spanning at least four types reduce $\mathrm{tr}(V_C)$ by 17% relative to sixteen controls from one dominant type; depth-validity statistics are given in Appendix E.

Figure 3 characterizes the operating regime of disjoint calibration. Fewer controls increase posterior uncertainty and move SCENECAL toward Marginal; type diversity contracts uncertainty more efficiently; and robust inference delays degradation as controls are contaminated. Together, these trends connect control informativeness directly to detection performance.

5.6 SHORTCUT CHECKS, TRANSFER, AND EFFICIENCY

Scene context carries dataset-level signal: controls-only AUC is 61.8/58.6/54.9% on WMCA/HQ-WMCA/HQ-MMFAS. However, replacing Protocol-B controls with condition-matched bona-fide donors changes ACER only from 4.72 to 4.96%, whereas random donors raise it to 7.31%. On the WMCA flexible-mask/glasses folds, controls-only AUC is 58.7/59.4% and IRLS downweights 7.8/2.5% of controls. The donor gap is consistent with acquisition matching, rather than native scene identity, being the useful component of context.

Table 3: Core ablation. Normalized bona-fide NLL and WMCA LOO ACER (%) ↓.

Variant	BF NLL ↓	LOO ACER ↓
SCENECAL	**1.00**	**2.99**
Type-agnostic response	1.14	4.27
Single global response	1.23	5.24
Concatenated modality features	1.11	4.08
CCA relation	1.01	3.06
Fixed target covariance	–	3.40
No held-out-control term	1.09	3.61
No face guard	1.12	4.24

Table 4: Cross-dataset RGB–Depth transfer, HTER (%) ↓. †: matched-protocol reimplementation.

Method	WMCA→HQ-WMCA	HQ-WMCA→WMCA
FM-ViT† (Liu et al., 2023)	20.2	10.0
Suppress & Rebalance† (2024) (Lin et al., 2024)	17.1	8.8
MMDG++† (Lin et al., 2025)	15.8	8.1
DADM† (Yang et al., 2025)	16.5	8.5
CTNet† (Chong et al., 2026)	15.3	7.8
CA-MMDG† (Zhang & Li, 2026)	16.9	8.7
Population	20.6	12.4
Marginal	18.1	9.8
Joint-plugin	16.5	8.1
SCENECAL	**14.7**	**7.3**

Table 4 reports cross-dataset transfer with source thresholds frozen. SCENECAL reaches 14.7/7.3% HTER for WMCA→HQ-WMCA and the reverse, compared with 15.3/7.8% for CTNet and 15.8/8.1% for MMDG++. Under severe NIR-gain, depth-sparsity, and exposure shifts on HQ-MMFAS, bona-fide BPCER increases by 1.8 points for SCENECAL, 3.6 for Marginal, and 5.4 for Population. These shifts also increase the norm of the inferred acquisition state, providing a direct diagnostic of source-to-target mismatch. End-to-end latency is 31.7 ms/frame for RGB–Depth and 44.1 ms/frame for RGB–Depth–NIR on one A100; the three-step robust calibration remains below 2 ms/frame (Appendix C).

6 Discussion and Conclusion

SCENECAL uses heterogeneous non-target regions as sample-specific evidence about acquisition state when target and nuisance effects are difficult to separate. Acquisition-state probes, condition-matched donor stability, and the control stress tests show that the inferred state follows capture changes more strongly than scene semantics, while reduced control information smoothly approaches the Marginal regime. The main applicability limits are tight crops, presentation media that occupy most of the scene, target-local sensor effects, and severe material shifts.

On public benchmarks, matched-protocol improvements over CTNet are modest but consistent under paired bootstrap; Benefits are also observed on held-out instruments, cross-dataset transfer, and sensor degradation. Supervised models remain competitive when attack families are known. Overall, the results support using disjoint scene evidence to condition verification when the target alone admits acquisition–attack ambiguity.

AI Use Statement

Generative AI tools assisted language editing; all technical claims and reported results were reviewed by the authors. The authors take full responsibility for the final content of this work.

Reproducibility Statement

We provide the information needed to reproduce the proposed method and the reported evaluations. Section 3 specifies the probabilistic model, relation construction, partial-observation handling, robust inference procedure, and training objective. Section 5 describes the datasets, modality settings, evaluation protocols, operating-point selection, and matched comparison settings. The appendix provides complete implementation details, preprocessing procedures, hyperparameters, runtime measurements, scene-validity statistics, null-calibration diagnostics, per-fold results, statistical uncertainty analyses, operating-point metrics, and additional ablations. The derivations of the analytical results are also included in the appendix. Public-dataset experiments follow the stated WMCA and HQ-WMCA protocols, and all preprocessing and calibration parameters are fitted only on the corresponding training splits.

Ethics Statement

This work studies face presentation attack detection and therefore involves biometric data. The public WMCA and HQ-WMCA datasets are used according to their respective research licenses and established evaluation protocols. HQ-MMFAS contains recordings collected from human participants. The data were collected with informed consent.

Our method additionally processes non-face regions of the captured scene, including clothing and background surfaces, because these regions provide information about acquisition conditions. Such context may contain incidental visual information unrelated to authentication. We therefore use scene context only to estimate acquisition variables and do not perform identity recognition, attribute inference, or semantic profiling from control regions. Results are reported only in aggregate.

HQ-MMFAS is not publicly released at the time of submission. Any future release will be subject to the consent, privacy, and data-use conditions under which the data were collected.

References

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Jun-Xiong Chong, Yu Hsu, Ming-Tsung Hsu, Yi-Ting Lin, Kai-Heng Chien, Chiou-Ting Hsu, and Pei-Kai Huang. Multi-modal face anti-spoofing via cross-modal feature transitions. *Expert Systems with Applications*, pp. 131292, 2026.

Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pp. 1704–1713. PMLR, 2018.

Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE transactions on information forensics and security*, 15:42–55, 2019.

A. Gijsenij, T. Gevers, and J. van de Weijer. Computational color constancy: Survey and experiments. *Trans. Img. Proc.*, 20(9):2475–2489, September 2011. ISSN 1057-7149. doi: 10.1109/TIP.2011.2118224. URL https://doi.org/10.1109/TIP.2011.2118224.

Guillaume Heusch, Anjith George, David Geissbühler, Zohreh Mostaani, and Sébastien Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):399–409, 2020.

Cheng-Yao Hong, Yen-Chi Hsu, and Tyng-Luh Liu. Contrastive learning for deepfake classification and localization via multi-label ranking. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17627–17637. IEEE, 2024.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. Fighting fake news: Image splice detection via learned self-consistency. In *ECCV*, 2018.

Binh M Le and Simon S Woo. Gradient alignment for cross-domain face anti-spoofing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 188–199. IEEE, 2024.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Wenzhong Tang, Zitong Yu, and Alex Kot. Suppress and rebalance: Towards generalized multi-modal face anti-spoofing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 211–221. IEEE, 2024.

Xun Lin, Ajian Liu, Zitong Yu, Rizhao Cai, Shuai Wang, Yi Yu, Jun Wan, Zhen Lei, Xiaochun Cao, and Alex Kot. Reliable and balanced transfer learning for generalized multimodal face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7608–7625, 2025.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383, 2010.

Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z. Li. 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.

Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang, Zhen Lei, Du Zhang, Stan Z Li, and Guodong Guo. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 18:4775–4786, 2023.

Owen Mayer and Matthew C Stamm. Exposing fake images with forensic similarity graphs. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1049–1064, 2020.

Zohreh Mostaani, Anjith George, Guillaume Heusch, David Geissbuhler, and Sebastien Marcel. The high-quality wide multi-channel attack (hq-wmca) database. Technical report, 2020.

Irving S Reed and Xiaoli Yu. Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution. *IEEE transactions on acoustics, speech, and signal processing*, 38 (10):1760–1770, 1990.

Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, and Wen-Sheng Chu. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24563–24574, 2023.

Di Tang, Zhe Zhou, Yinqian Zhang, and Kehuan Zhang. Face flashing: a secure liveness detection protocol based on light reflections. 2018.

Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20281–20290, June 2022.

Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, and Kui Ren. Fsfm: A generalizable face security foundation model via self-supervised facial representation learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24364–24376. IEEE, 2025.

Jingyi Yang, Xun Lin, Zitong Yu, Liepiao Zhang, Xin Liu, Hui Li, Xiaochen Yuan, and Xiaochun Cao. Dadm: Dual alignment of domain and modality for face anti-spoofing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12045–12056. IEEE, 2025.

Zitong Yu, Ajian Liu, Chenxu Zhao, Kevin HM Cheng, Xu Cheng, and Guoying Zhao. Flexible-modal face anti-spoofing: A benchmark. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6346–6351. IEEE, 2023a.

Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b. doi: 10.1109/TPAMI.2022.3215850.

Botao Zhang and Hongjiao Li. Domain generalization for multi-modal face anti-spoofing via mixcropmask augmentation and deep hybrid-level fusion. *Engineering Applications of Artificial Intelligence*, 158:111287, 2025.

Botao Zhang and Hongjiao Li. Contrast and adversarial: Towards generalized multi-modal face anti-spoofing. *Engineering Applications of Artificial Intelligence*, 176:114783, 2026.

Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 919–928. IEEE, 2019.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15023–15033, October 2021.

Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma. Instance-aware domain generalization for face anti-spoofing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20453–20463. IEEE, 2023.

Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, and Lizhuang Ma. Test-time domain generalization for face anti-spoofing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 175–187. IEEE, 2024.