DETECTING AUDIO-VISUAL DEEPFAKES OUTSIDE THE GENERATOR OBJECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

A generator may match an audio–visual synchronization objective at an observed sample while retaining a different local articulatory response. We propose **Outside the Objective (OtO)**, an authentic-only test of how acoustic features inferred by a visual-to-acoustic predictor change under local mouth-tangent queries. Our central construction maps declared generator objectives into the acoustic response space and tests the complementary response conditioned on the observed audio. A canonical detector-space cosine gives an exact local plane; declared differentiable losses contribute Riesz representatives. Conditional learning and objective projection use one consistent residual, with a compact probe budget for efficient evaluation. In a controlled synchronization-weight sweep, the zero-order cosine detector falls from 88.9 to 66.9 AUROC while the projected first-order test remains between 90.2 and 90.7. Under a common cross-domain protocol, OTO reaches 92.0% AUROC on MMDF and 84.1% on HiFi-AVDF, with 3.6 and 4.8 points attributable to objective-direction removal beyond audio conditioning.

1 INTRODUCTION

Audio–visual deepfake detection has increasingly moved from pixel artifacts toward cross-modal correspondence. Authentic-only methods model normal audio–visual dynamics or speech representations (Feng et al., 2023; Liang et al., 2024; Smeu et al., 2025), while recent systems learn stronger alignment, pseudo-fake, and multimodal fusion cues (Du et al., 2025; Shahzad et al., 2026; Wei and Li, 2026; Peng et al., 2026; Kim et al., 2026). These directions improve transfer, but they share a structural vulnerability: correspondence itself can be a training target of the generator. Lip-sync objectives built from SyncNet-like experts (Prajwal et al., 2020) and audio–visual speech objectives such as AV-HuBERT (Shi et al., 2022) explicitly reward the same zero-order agreement that many detectors measure. Empirically, cross-modal correlation is not monotone in authenticity across manipulation families (Yu et al., 2024); a highly optimized fake can therefore look more "synchronized" to an alignment score without reproducing the local response law of genuine speech.

This motivates an objective-relative question. Rather than asking whether the observed mouth and audio agree under a score the generator may already optimize, we ask what remains *outside the generator objective*: after mapping declared training objectives into a common observable response space and removing those directions, does a local articulatory change still produce the acoustic response expected from authentic speech, conditioned on the actual audio? This shifts the evidence from zero-order alignment to first-order response geometry. Our theory characterizes this regime under local objective matching, and the controlled intervention measures how closely trained generators approach it as objective strength increases.

We use a real-only visual-to-acoustic predictor and local mouth-mesh tangents as a model-based measurement interface. Symmetric tangent queries measure the predictor's local acoustic sensitivity to controlled mouth-trajectory changes. A canonical cosine alignment yields an exact sample-dependent plane in the detector space, while declared generator losses are pulled back separately as observable Riesz representatives. An authentic audio-conditioned reference models the expected response given content, prosody, and pose, and the final statistic uses only the conditional residual orthogonal to the objective span. Probe selection is applied afterward to reduce evaluation cost.

Our contributions are threefold: (i) objective-relative acoustic response geometry with an exact local cosine plane and observable pullbacks for differentiable objectives; (ii) a local relation characterizing

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.

when objective matching constrains objective-aligned first-order response; and (iii) an authentic-only conditional residual test that preserves the objective complement from design through scoring. Controlled interventions and shortcut audits validate these mechanisms.

2 Related Work

Audio–visual deepfake detection. Authentic-only approaches model normal audio–visual dynamics or speech representations without labeled fakes, including AVAD, SpeechForensics, and AVH-Align (Feng et al., 2023; Liang et al., 2024; Smeu et al., 2025). Recent 2025–2026 systems instead emphasize cross-modal alignment, pseudo-fakes, or multimodal fusion, including CAD, SAVe, AVPF, MSCT, HAVIC, and X-AVDT (Du et al., 2025; Shahzad et al., 2026; Wei and Li, 2026; Wei et al., 2026; Peng et al., 2026; Kim et al., 2026). Yu et al. (2024) shows that synchronization/correlation itself is not a uniform cue across manipulation families. Our distinction is orthogonal to the training regime: we test a first-order conditional response after removing directions induced by generator objectives, rather than scoring another zero-order correspondence.

Matched-subspace detection motivates rejecting specified interference directions (Scharf and Friedlander, 1994; Kraut and Scharf, 1999). We derive a sample-dependent cosine plane and observable directions induced by generator losses using frozen speech representations and synchronization experts (Chen et al., 2022; Hsu et al., 2021; Prajwal et al., 2020; Shi et al., 2022). Architectural coupling (Park et al., 2022) further motivates distinguishing representable from unreported objectives.

3 Objective-Relative Acoustic Responses

3.1 Local articulatory response

Let v be a tracked one-second mouth-mesh sequence, a its aligned observed audio, and $f_\theta(v) = \hat{a} \in \mathbb{R}^d$ a frozen predictor trained only on authentic pairs. Let $\mathcal{Q} = \{q_m\}_{m=1}^M$ consist of unit-norm local mesh-velocity tangents. A valid perturbation obeys $T_{tq}v = v + tq + O(t^2)$ in local coordinates; the symmetric response is

$$d_q(v;t) = \frac{f_\theta(T_{tq}v) - f_\theta(T_{-tq}v)}{2t} = J_f(v)q + O(t^2). \tag{1}$$

The finite-difference remainder is $O(t^2)$ for a sufficiently smooth composed trajectory. Equation (1) defines a model-based measurement interface: d_q records the sensitivity of the frozen predictor f_θ to a controlled mouth-trajectory perturbation. We interpret this quantity as a predictor response rather than a directly observed physical response, and Section 5.5 tests whether the resulting evidence is stable across heterogeneous predictors. Because f_θ consumes the tracked mesh sequence only, the query stays within its geometric input domain. Authentic context c indexes coarse speech content and pose. Define $m_{c,q} = \mathbb{E}_0[d_q \mid c]$ and $\Gamma_0 = \mathrm{Cov}_{(x,q)\sim P_0\times \mathrm{Unif}(\mathcal{Q})}[d_q(x) - m_{c(x),q}]$. With $W = (\Gamma_0 + \gamma I)^{-1/2}$ fixed on an authentic fit split, $\bar{d}_q = W d_q$ is the response in the metric used for objective projection.

3.2 Objectives in the observable space

For the canonical detector-space cosine $c_{\mathrm{can}}(v,a) = u^\top e$, where $u = \hat{a}/\|\hat{a}\|$, $e = a/\|a\|$ and $\tau = u^\top e$, differentiation gives

$$\dot{c}_q = \frac{d_q^\top (e - \tau u)}{\|\hat{a}\|}. \tag{2}$$

Since $d_q = W^{-1}\bar{d}_q$, the Euclidean Riesz vectors of the cosine covector and its radial normalization direction in whitened coordinates belong to

$$\mathcal{S}_{\cos}(x) = \mathrm{span}\{W^{-1}u, W^{-1}e\}. \tag{3}$$

Thresholded SVD yields its rank-one or rank-two basis directly, including near collinearity. This plane is exact for c_{can}; it is not identified with a SyncNet or AV-HuBERT training loss.

For a declared differentiable objective, write $\widetilde{\ell}_j(v,a) = \ell_j(\mathcal{R}_x(v), a)$ when the expert consumes pixels, where $\mathcal{R}_x$ is a fixed mouth-region warp driven by the same mesh displacement. Let $b_j(x)_m =$

$\partial_t \widetilde{\ell}_j(T_{tq_m} v, a)\Big|_0$ and $\bar{D}_x = [\bar{d}_{q_1}, \ldots, \bar{d}_{q_M}]$; derivatives use autodiff through the composed expert when available and symmetric finite differences otherwise. Its observable regularized representative is

$$s_j(x) = \bar{D}_x(\bar{D}_x^\top \bar{D}_x + \rho I)^{-1} b_j(x). \tag{4}$$

We set $\rho = 10^{-3}$ times the median nonzero eigenvalue of $\bar{D}_x^\top \bar{D}_x$, with a frozen authentic-validation sweep of 10^{-4}–10^{-1}. Context-indexed $B_{\mathrm{emp}}(c)$ retains reproducible directions of s_j across disjoint authentic folds. All columns below are expressed in the same whitened response coordinates: $W^{-1}u$ and $W^{-1}e$ are the Riesz vectors of the original cosine covectors, and $B_{\mathrm{emp}}(c)$ is learned from $\bar{D}_x$. With $U_{\mathcal{F}}(x) = \mathrm{orth}[W^{-1}u, W^{-1}e, B_{\mathrm{emp}}(c)]$, project the complete candidate matrix:

$$R_x = (I - U_{\mathcal{F}}(x)U_{\mathcal{F}}(x)^\top)\bar{D}_x. \tag{5}$$

The canonical cosine plane is exact for each sample; the empirical basis is a frozen context approximation. Its held-out representative capture is measured by $\|\Pi_{B_{\mathrm{emp}}(c)} s_j(x)\|/\|s_j(x)\|$.

Objective coverage. Differentiable expert losses enter through Equation (4), while observable architectural or discrete constraints use directional finite differences. Each represented loss is retained only when its signature magnitude $\|b_j\|$ and reconstruction error $\|\bar{D}_x^\top s_j - b_j\|/(\|b_j\| + 10^{-8})$ pass the authentic-only checks in the appendix; unreported objectives are evaluated separately in Section 5.7.

3.3 Local objective-relative response principle

For a fixed full-rank context basis $B = [s_1, \ldots, s_J]$, let Δ be the conditional authentic-to-generated mean shift, $g = B^\top \Delta$, and $\Pi_B = \mathrm{Proj}(\mathrm{span}(B))$.

Proposition 1 (Local suppression under objective matching)**.** *Under the local identity-covariance mean-shift model,*

$$\|\Delta\|^2 = \|\Pi_B \Delta\|^2 + \|(I - \Pi_B)\Delta\|^2, \qquad \|\Pi_B \Delta\|^2 \le \frac{\|g\|^2}{\sigma_{\min}^2(B)}. \tag{6}$$

For an objective j, let $h_j(t) = \mathbb{E}_G[\ell_j(T_{tq}X)] - \mathbb{E}_0[\ell_j(T_{tq}X)]$ compare samples with a matched audio/context distribution. If $|h_j(\pm t)| \le \varepsilon_j(t)$, $|h_j''(s)| \le L_j$ for $|s| \le t$, and the fixed-basis derivative pullback has discrepancy at most δ_j, then

$$|g_j| \le \inf_{0<t\le r} \left\{ \frac{\varepsilon_j(t)}{t} + \frac{L_j t}{2} \right\} + \delta_j. \tag{7}$$

When the local matching conditions in Proposition 1 hold, the represented objective component is constrained. The condition is stronger than the mere inclusion of ℓ_j in a generator training loss; Section 5.2 measures objective satisfaction and aligned response as the loss strength varies.

Role of projection. Projection is an inductive bias rather than a source of new authenticity information. For fixed conditioning variables and a known Π_x, the projected response is a deterministic restriction of the full residual, so an unrestricted oracle detector can emulate any projected decision rule. OTO instead restricts finite-sample estimation away from directions whose distribution changes with generator-objective optimization, targeting more stable cross-generator evidence.

4 Method

Figure 1 summarizes the pipeline: local mesh queries define acoustic responses, generator objectives define the removed directions, and observed audio conditions the complementary response used for design and scoring.

4.1 An objective-compatible conditional reference

Our frozen authentic visual-to-acoustic predictor outputs $d = 768$-dimensional WavLM features. Frozen HuBERT speech units and plausible mesh velocities define $M = 64$ unit candidate tangents.

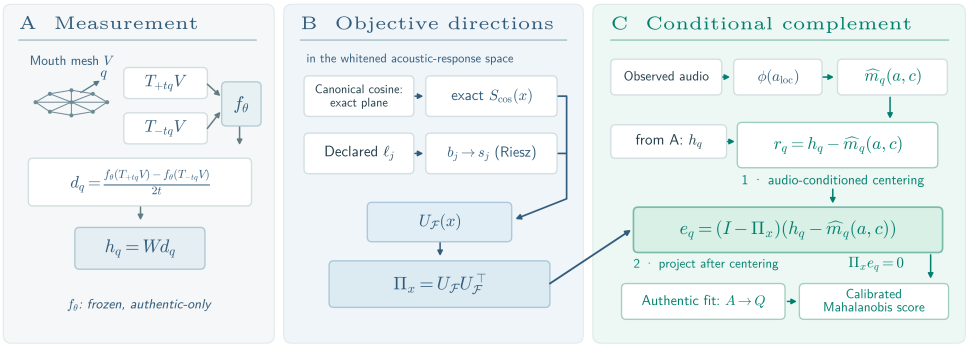

Figure 1: Overview of OTO. Local mesh queries yield whitened acoustic responses; generator objectives define the removed subspace; observed audio conditions the complement used for probe design and calibrated scoring.

Let $h_q(x) = \bar{d}_q(x)$ denote the whitened *unprojected* response from Equation (1). The observed audio defines a $p = 32$-dimensional context vector $\phi(a_{\text{loc}})$. The main configuration uses a frozen speech-content/prosody bottleneck (unit posteriors, pitch and energy); compressed WavLM is a matched-capacity comparison. The coarse index c includes speech and pose.

On disjoint authentic speakers, a low-rank shared basis $V \in \mathbb{R}^{d \times L_0}$, $L_0 = 32$, defines an unprojected conditional reference for *all* candidate probes:

$$\widehat{m}_q(a,c) = V\big(\beta_{c,q} + G_{c,q}\phi(a_{\text{loc}})\big), \quad \beta_{c,q} \in \mathbb{R}^{L_0}, \quad G_{c,q} \in \mathbb{R}^{L_0 \times p}. \tag{8}$$

A pooled ridge prior stabilizes sparse context heads; hyperparameters and V are selected with authentic-only held-out speakers. For each input define $\Pi_x = U_{\mathcal{F}}(x)U_{\mathcal{F}}(x)^\top$ using Equation (3) and the frozen context objective basis. Let $r_q(x) = h_q(x) - \widehat{m}_q(a,c)$ be the audio-conditioned residual. The *operative* response is

$$\boxed{e_q(x) = (I - \Pi_x)r_q(x)}, \qquad \Pi_x e_q(x) = 0. \tag{9}$$

Projection is performed after conditional centering; the PCA basis and probe design are then learned only from e_q, so removed coordinates cannot re-enter through compression or selection.

4.2 Response basis and probe budget

We learn the orthonormal $A \in \mathbb{R}^{d \times L}$, $L = 8$, from authentic out-of-fold e_q and set $z_q = A^\top e_q \in \mathbb{R}^L$. The objective-relative test is defined for all $M = 64$ candidate responses. To reduce inference cost, we select $K = 16$ probes only after Equation (9), using an authentic-only noise-adjusted E-optimal criterion. The selection objective, rounding procedure, and budget ablations are given in Section C; they affect computational coverage but not the definition of the projected residual.

4.3 Score, unit of analysis, and calibration

Stack $z_x = \text{vec}([A^\top e_q(x)]_{q \in Q}) \in \mathbb{R}^{KL}$; estimate an authentic shared shrunk covariance Σ_* with pooled scalar scales $s_{c,r,b}$ indexed by context, cosine-plane rank r, and valid-window-count bin b. For a valid window,

$$S_{\text{win}}(x) = z_x^\top (s_{c,r,b}\Sigma_*)^{-1} z_x. \tag{10}$$

A clip score is the median of its accepted window scores. Tracking/SNR selects valid windows without using authenticity scores; a clip with no valid track is a nondecision. The *primary analysis unit* throughout the main tables is one eligible source-linked clip; all methods use identical deterministic source-block sampling and fixed aggregation. Speakers and source videos form bootstrap and train/calibration/test separation blocks. Speaker-aggregated results are an additional analysis, not substituted into the video/clip ROC.

Table 1: Evaluation composition in independent source files. Appearance-only files are evaluated separately from the direct-mechanism pool.

Dataset	Real	Visual-artic. fake	Audio fake	Joint fake	Appearance-only
AV-DF1M++	6000	5760	1440	4800	600
DeepSpeak	1800	630	140	630	180
MMDF	2000	2900	400	1700	500
HiFi-AVDF	1800	1960	520	1520	400

On independent source-authentic calibration units, each predeclared group g has a split-conformal upper quantile q_g at $\alpha = 0.05$. For overlapping groups,

$$q(x) = \max_{g \in \mathcal{G}(x)} q_g, \qquad \widehat{y}(x) = \mathbf{1}\{S(x) > q(x)\}. \tag{11}$$

The group basis covers pose/yaw, blur, illumination, compression, language, speech activity and valid-window-count bins, with sparse strata pooled before threshold fitting. Calibration is frozen for all targets. Within-group marginal control refers to exchangeable independent source units under the same fixed selection rule; realized finite-test FPR and uncertainty are reported separately.

5 Experiments

5.1 Protocol, Data Splits, and Edit Taxonomy

All pre-score components use authentic data only, with 7,000/3,500/3,500/6,000 files for predictor/reference fitting, residual/probe design, threshold calibration, and source testing. Splits are identity/source-video disjoint, all design choices are frozen before calibration, and bootstrap resampling is paired by source and clustered by identity.

All methods use the same source-linked clips in Table 1. Supervised or pseudo-fake baselines train only on the AV-DF1M++ source split, while label-free methods see source-authentic data only. We use source-block bootstrap for AUROC comparisons and a common source-authentic 5%-FPR calibration for threshold transfer; target thresholds appear only in the oracle analysis.

We evaluate AV-Deepfake1M++, DeepSpeak, MMDF, and HiFi-AVDF (Cai et al., 2025; Barrington et al., 2026; Kim et al., 2026; Peng et al., 2026). A controlled suite with declared SyncNet, AVSyncNet, and AV-HuBERT objectives supports mechanism interventions. Visual-articulation, audio, and joint edits form the direct-mechanism pool, while appearance-only identity swaps are evaluated separately because they need not alter articulatory–acoustic coupling.

For a shared authentic distribution R and fake mixture $F = \sum_k w_k F_k$,

$$\mathrm{AUROC}(F, R) = \sum_k w_k \, \mathrm{AUROC}(F_k, R), \tag{12}$$

up to ties. Category-specific analyses therefore share the same authentic pool and are checked against the pooled score using the fixed counts in Table 1.

The primary metrics are clip AUROC, normalized pAUC for FPR $\leq 10\%$, and TPR at 1% and 5% FPR. Headline differences use 10,000-resample paired source-block bootstrap.

5.2 Increasing objective strength suppresses aligned evidence

Figure 2 directly intervenes on the generator objective. We vary only λ_{sync} while holding the architecture, source videos, training schedule, and visual-quality bins fixed. The same 1,800 sources and three generator seeds are used at every intervention value.

Increasing synchronization optimization raises external sync quality from 0.60 to 0.84 and reduces the normalized objective-aligned response from 0.143 to 0.028. Over the same trajectory, zero-order cosine AUROC falls from 88.9 to 66.9 and the unprojected first-order score falls to 84.6. In contrast, OTO remains between 90.2 and 90.7 AUROC. Paired bootstrap half-widths across the sweep range

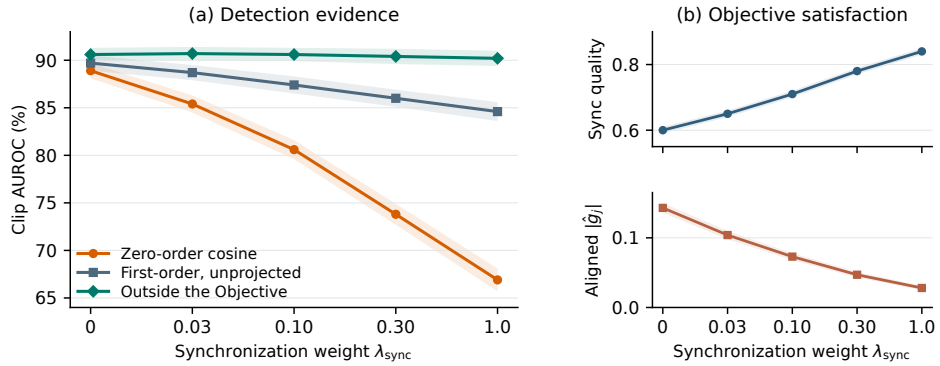

Figure 2: Controlled synchronization-objective intervention. Stronger synchronization optimization suppresses zero-order and objective-aligned evidence, while the objective-complement first-order score remains stable. Error ribbons use paired source-block bootstrap.

Table 2: Matched-objective interventions measured by clip AUROC (%). Bold entries remove the objective used to train the corresponding generator.

Generation objective	None	SyncNet	AVSyncNet	AV-HuBERT	Random	Energy-match	Union
SyncNet ($\lambda = 0.10$)	87.4	**90.6**	86.9	86.7	87.6	87.9	90.0
AVSyncNet	88.8	88.1	**92.0**	87.9	89.0	89.3	91.2
AV-HuBERT	89.2	88.4	88.3	**92.3**	89.5	89.7	91.6

from 0.7 to 1.2 points. Together, these curves show that, in this controlled family, increasing objective strength both improves objective satisfaction and suppresses aligned evidence, consistent with the suppression pattern characterized by Proposition 1 while the complementary first-order score remains stable.

5.3 The removed direction is objective specific

Table 2 tests whether the gain comes from removing the generator's objective direction rather than from generic dimension reduction. Each controlled generator is evaluated with the same conditional pipeline while only the removed subspace is changed. Same-rank random, authentic-energy-matched, and mismatched experts provide controls with comparable dimensionality or removed variance.

Removing the matched objective improves AUROC by 3.2, 3.2, and 3.1 points for SyncNet, AVSyncNet, and AV-HuBERT generators, respectively. Their paired 95% confidence intervals are $[+2.3, +4.1]$, $[+2.2, +4.2]$, and $[+2.1, +4.1]$. Energy-matched removal improves each row by only about 0.5 point, showing that the benefit is objective-specific rather than a generic consequence of deleting high-variance directions. The union basis remains within 0.6–0.8 point of the corresponding matched expert while covering all declared objectives.

5.4 Projection as a finite-sample and objective-shift bias

We compare no, soft, and hard objective suppression through $e_q^{(\alpha)} = (I - \alpha\Pi_x)r_q$ with $\alpha \in \{0, 0.5, 1\}$. The finite-sample sweep varies only the authentic design set, while the mismatch sweep rotates the removal basis toward an equal-rank nonmatching basis with the generator and evaluation set fixed.

Table 3 separates finite-sample and mismatch effects. From 500 to 3,500 design files, the hard-projection gain over no suppression narrows from 4.7 to 3.1 points but remains positive. Under basis mismatch, hard projection is strongest near the matched objective, whereas soft suppression degrades more gradually and becomes preferable at $60°$. These trends indicate contributions from both finite-sample regularization and objective-shift robustness.

Table 3: Projection-bias audit on the controlled generator suite, measured by mean clip AUROC (%). Soft uses $\alpha = 0.5$ and hard uses $\alpha = 1$. The sample-size panel varies only authentic design data; the mismatch panel rotates the removal basis while keeping its rank fixed.

Setting	None	Soft	Hard
Authentic design-set size			
500	84.6	87.8	89.3
1,000	86.1	89.0	90.4
2,000	87.4	90.0	91.1
3,500	88.5	90.7	91.6
Objective-basis mismatch			
0°	88.5	90.7	91.6
20°	88.5	90.6	91.3
40°	88.5	90.2	90.4
60°	88.5	89.6	89.1

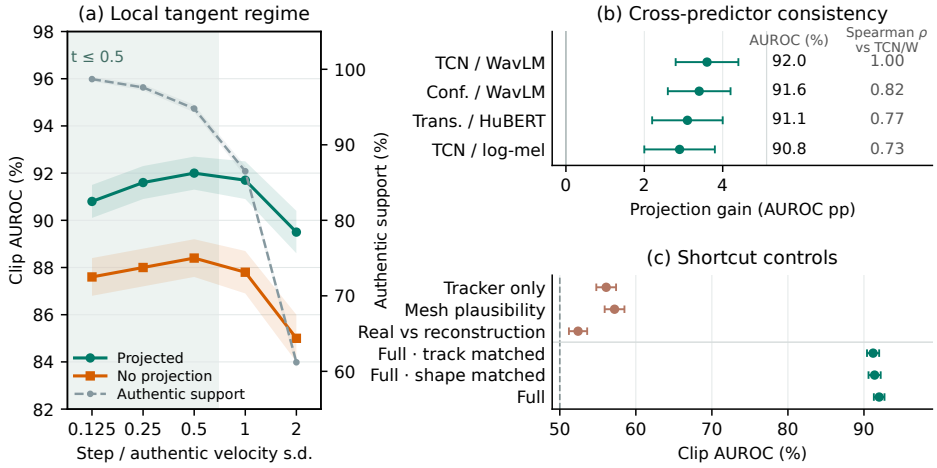

Figure 3: Competing-explanation audit. The gain appears within the authentic local-support regime, persists across heterogeneous predictors, and remains after matching tracking and mesh-plausibility quality. Shortcut-only controls remain near chance.

5.5 Competing Explanations

Figure 3 tests local-support, predictor-specific, and tracking/trajectory shortcut explanations.

The projection gain is already present for $t/\sigma_v \leq 0.5$, where at least 94.8% of perturbed trajectories remain inside the authentic local-support envelope. Across three additional heterogeneous predictors, score correlations with the default instrument are 0.82, 0.77, and 0.73, while full AUROC remains 91.6, 91.1, and 90.8. Tracker-only, mesh-plausibility-only, and real-versus-reconstruction controls reach 56.1, 57.2, and 52.4 AUROC. After matching tracking quality and mesh plausibility, the full detector retains 91.2 and 91.4 AUROC, compared with 92.0 before matching.

The same frozen score is also applied to three external authentic corpora before fake evaluation. Their FPRs range from 6.8% to 8.1%, while source-real versus target-real score AUROC ranges from 0.58 to 0.61. Refitting the authentic reference on target-domain real data changes fake AUROC by only +0.3, +0.6, and +0.8 point on DeepSpeak, MMDF, and HiFi-AVDF. Codec and resolution matching changes the MMDF and HiFi-AVDF results by -0.7 and -1.0 point.

5.6 First-order evidence, audio conditioning, and modality attribution

The same-backbone zero-order control reuses f_θ, whitening, the audio bottleneck, score dimension, and calibration, changing only the order of the response statistic. It obtains 85.6% and 73.9% AUROC on MMDF and HiFi-AVDF, compared with 92.0% and 84.1% for the first-order test. The paired gains are $+6.4$ $[5.2, 7.6]$ and $+10.2$ $[8.7, 11.7]$ points.

Table 4: Factorial attribution measured by clip AUROC (%) for audio conditioning and objective projection.

Configuration	MMDF	HiFi-AVDF
Coarse context, no projection	86.9	75.5
Coarse context + objective projection	89.6	79.5
Continuous audio, no projection	88.4	79.3
Continuous audio + objective projection	**92.0**	**84.1**

Table 5: Modality diagnostics measured by clip AUROC (%). Appearance-only identity swaps are evaluated as a specificity control and are excluded from the direct-mechanism pool.

Dataset	Configuration	Visual articulation	Audio	Joint	Appearance-only
MMDF	Full content/prosody test	94.4	71.0	92.8	58.3
MMDF	ϕ-only	50.8	52.1	51.5	50.6
MMDF	Full compressed-WavLM	94.8	76.8	93.1	59.0
MMDF	ϕ-only compressed-WavLM	52.3	64.0	58.1	51.0
HiFi	Full content/prosody test	88.6	65.4	84.7	56.8
HiFi	ϕ-only	51.2	52.6	51.9	50.8
HiFi	Full compressed-WavLM	89.0	70.6	85.0	57.4
HiFi	ϕ-only compressed-WavLM	52.1	60.5	55.7	50.9

Table 4 separates continuous audio conditioning from objective projection in a 2×2 factorial design. Within the continuous-audio condition, objective projection improves AUROC by 3.6 points $[2.8, 4.4]$ on MMDF and 4.8 points $[3.6, 6.0]$ on HiFi-AVDF. Within the projected test, continuous conditioning contributes 2.4 and 4.6 points, respectively. The two factors therefore make distinct contributions, while their combination gives the strongest result.

Table 5 separates manipulation modalities. The bottleneck alone remains near chance, whereas the full response test is strongest on visual-articulation and joint edits; audio-only edits are harder and appearance-only swaps remain close to chance. Using the fixed counts in Table 1, the category results recover the pooled 92.0/84.1 AUROC through Equation (12).

Natural synchronization tolerance. On 1,800 authentic sources, the frozen threshold gives 5.0% FPR at zero offset, 7.0% at $+45$ ms, 6.1% at -60 ms, and 8.6% at -125 ms. FPR increases to 10.9% at $+90$ ms, 11.7% at -185 ms, and 47.0% at $+200$ ms. Same-speaker different-content swaps reach 88.0% AUROC, whereas content- and timing-preserving voice cloning remains near 55.0%.

5.7 Objective scope and adaptive generation

The projection gain is about 3.1 points for generators with documented objectives. It remains positive for unreported-objective strata, reaching $+1.6$ $[0.7, 2.5]$ on MMDF and $+2.1$ $[1.0, 3.2]$ on HiFi-AVDF. Across the controlled declared experts, 94.0% of windows yield observable objective signatures. The median Riesz reconstruction error is 0.08 and held-out empirical-basis capture is 0.83.

Table 6 evaluates generators adapted with black-box, gray-box, and white-box access. The white-box attack directly minimizes the final clip score $S(x)$. Performance decreases most under full white-box access, from 90.6 to 76.1 AUROC. Once the newly exposed attack loss is represented in the objective library, AUROC recovers to 87.9 without fake-label retraining. The corresponding recovery is smaller for black- and gray-box attacks because their pre-update degradation is already more limited.

5.8 Operating-point and authentic-reference transfer

Table 7 evaluates a common source-authentic 5%-FPR threshold; oracle columns re-calibrate on target authentic data. OTO transfers to 6.7/74.0 FPR/TPR on DeepSpeak and 8.4/67.4 on HiFi-AVDF, with target-real KS distances of 0.083 and 0.107.

English/Spanish/Mandarin bins yield 5.0/5.5/6.0% FPR, while compression, resolution, blur, low light, and sensor noise shift FPR by at most 2.5 points. Full subgroup statistics are reported in the appendix.

Table 6: Adaptive-generator audit measured by AUROC (%). The post-update setting adds the revealed attack loss to the objective library without fake-label retraining.

Attacker knowledge	Before attack	After attack	After objective update
Black-box	90.6	88.2	89.8
Gray-box	90.6	84.6	89.1
White-box	90.6	76.1	87.9

Table 7: Target operating points at clip level (%). Oracle TPR uses a target-authentic 5%-FPR threshold, while transferred results retain the source-authentic threshold.

	DeepSpeak			HiFi-AVDF		
Method	FPR	TPR	Oracle TPR	FPR	TPR	Oracle TPR
AVH-Align	9.0	66.8	62.4	13.4	57.0	50.6
CAD	12.1	72.8	64.0	17.8	62.0	52.1
SAVe	9.5	69.5	64.9	14.0	63.3	55.7
AVPF	8.9	71.2	67.0	12.8	65.1	58.5
HAVIC	10.5	73.0	66.1	16.2	61.5	52.9
X-AVDT	10.8	72.9	65.8	14.3	65.7	59.7
OTO	**6.7**	**74.0**	**71.0**	**8.4**	**67.4**	**63.1**

Table 8: Common-protocol clip AUROC (%). LF/PF denote no labeled/pseudo-fake training.

Method	Year	LF	PF	AV-DF1M++	DeepSpeak	MMDF	HiFi-AVDF
AVAD (Feng et al., 2023)	2023	✓	✓	78.5	73.1	74.9	67.0
SpeechForensics (Liang et al., 2024)	2024	✓	✓	83.0	79.2	80.2	71.8
Tian et al. (Tian et al., 2025)	2025	✓	✓	79.0	76.1	77.5	69.4
AVH-Align (Smeu et al., 2025)	2025	✓	✓	84.7	80.4	81.2	72.3
CAD (Du et al., 2025)	2025	×	×	89.3	86.0	86.8	77.1
MSCT (Wei et al., 2026)	2026	×	×	88.8	85.3	86.2	76.7
SAVe (Shahzad et al., 2026)	2026	✓	×	88.0	84.1	85.5	76.2
AVPF (Wei and Li, 2026)	2026	✓	×	89.7	85.5	86.5	76.7
HAVIC (Peng et al., 2026)	2026	×	×	91.2	88.0	88.5	76.4
X-AVDT (Kim et al., 2026)	2026	×	×	91.6	87.9	90.4	79.8
OTO	–	✓	✓	90.8	**89.8**	**92.0**	**84.1**

5.9 Comparison with recent methods

Table 8 compares all methods under the common protocol. Relative to X-AVDT, OTO changes AUROC by -0.8 $[-1.5, -0.1]$ on AV-DF1M++ and by $+1.9$ $[0.7, 3.1]$, $+1.6$ $[0.6, 2.6]$, and $+4.3$ $[3.0, 5.6]$ on DeepSpeak, MMDF, and HiFi-AVDF.

Runtime is 0.68 s per clip at $K = 16, L = 8$; the $K = 64, L = 32$ setting reaches 93.1/85.0 AUROC on MMDF/HiFi-AVDF at 2.42 s. Full sweeps are in the appendix.

6 Discussion and Conclusion

The results support objective-relative detection as an inductive bias rather than an information-creating operation: optimized objective directions become less reliable, while their complementary first-order response remains stable across finite-sample, objective-shift, and cross-domain tests. OTO therefore targets manipulations that alter the articulatory–acoustic conditional law; appearance-only edits, content/timing-preserving voice cloning, strong white-box adaptation, and large capture desynchronization remain important boundary cases. Reproducibility artifacts and complete diagnostic results are provided in the supplementary material.

AI USE STATEMENT

Generative AI tools assisted language editing; all technical claims and reported results were reviewed by the authors.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the method and evaluation across the main paper and appendices. Section 5.1 specifies the data isolation, source-linked analysis unit, calibration protocol, evaluation taxonomy, and bootstrap procedure. Appendix A gives the complete proof of Proposition 1. Appendix B specifies the visual-to-acoustic predictor, tangent construction, objective pullback implementation, empirical objective-basis construction, hyperparameters, sample sizes, subgroup denominators, and baseline provenance. Appendix C reports probe and response-budget sweeps, alternative scoring families, and robustness results. The paper additionally reports the randomization and grouping units used for statistical uncertainty so that confidence intervals can be reproduced from source-level predictions. The supplementary material includes the source-linked manifests, objective/expert configurations, selected probe indices, subgroup definitions, and evaluation specifications used in the reported experiments.

ETHICS STATEMENT

This work studies audio–visual deepfake detection using existing research benchmarks and does not conduct new human-subject experiments or collect new personal data. We use the benchmark data only for the research purposes and evaluation settings described in this paper and follow the corresponding dataset licenses and usage conditions.

Deepfake detection can support media forensics and content-integrity analysis, but detector outputs can also be misinterpreted as definitive judgments of authenticity. Our experiments show that the proposed method has identifiable boundary cases, including appearance-only manipulations, content- and timing-preserving voice cloning, strong white-box adaptation, and substantial audio–video desynchronization. We therefore do not intend the detector score to serve as the sole basis for consequential decisions about individuals or media provenance. In practical deployment, detector outputs should be combined with additional forensic evidence and appropriate human review.

REFERENCES

Sarah Barrington, Maty Bohacek, and Hany Farid. The deepspeak dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1893–1902, 2026.

Zhixi Cai, Kartik Kuckreja, Shreya Ghosh, Akanksha Chuchra, Muhammad Haris Khan, Usman Tariq, Tom Gedeon, and Abhinav Dhall. Av-deepfake1m++: A large-scale audio-visual deepfake benchmark with real-world perturbations, 2025.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

Yuxuan Du, Zhendong Wang, Yuhao Luo, Caiyong Piao, Zhiyuan Yan, Hao Li, and Li Yuan. Cad: A general multimodal framework for video deepfake detection via cross-modal alignment and distillation, 2025.

Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10491–10503. IEEE, 2023.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Youngseo Kim, Kwan Yun, Seokhyeon Hong, Sihun Cha, Colette Suhjung Koo, and Junyong Noh. X-avdt: Audio-visual cross-attention for robust deepfake detection, 2026.

Shawn Kraut and Louis L Scharf. The cfar adaptive subspace detector is a scale-invariant glrt. *IEEE Transactions on Signal Processing*, 47(9):2538–2541, 1999.

Yachao Liang, Min Yu, Gang Li, Jianguo Jiang, Boquan Li, Feng Yu, Ning Zhang, Xiang Meng, and Weiqing Huang. Speechforensics: Audio-visual speech representation learning for face forgery detection. volume 37, pages 86124–86144, 2024.

Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2062–2070, 2022.

Jielun Peng, Yabin Wang, Yaqi Li, Long Kong, and Xiaopeng Hong. Leave no stone unturned: Uncovering holistic audio-visual intrinsic coherence for deepfake detection. 2026.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild, 2020.

Louis L Scharf and Benjamin Friedlander. Matched subspace detectors. *IEEE Transactions on signal processing*, 42(8):2146–2157, 1994.

Sahibzada Adil Shahzad, Ammarah Hashmi, Junichi Yamagishi, Yusuke Yasuda, Yu Tsao, Chia-Wen Lin, Yan-Tsung Peng, and Hsin-Min Wang. Save: Self-supervised audio-visual deepfake detection exploiting visual artifacts and audio-visual misalignment, 2026.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. 2022.

Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18815–18825. IEEE, 2025.

Mulin Tian, Mahyar Khayatkhoei, Joe Mathai, and Wael AbdAlmageed. Unsupervised multimodal deepfake detection through explicit intra-modal and cross modal inconsistency discovery. 2025.

Fangda Wei, Miao Liu, Yingxue Wang, Jing Wang, Shenghui Zhao, and Nan Li. Msct: Differential cross-modal attention for deepfake detection, 2026.

Zihe Wei and Yuezun Li. Generalizing video deepfake detection by self-generated audio-visual pseudo-fakes, 2026.

Cai Yu, Shan Jia, Xiaomeng Fu, Jin Liu, Jiahe Tian, Jiao Dai, Xi Wang, Siwei Lyu, and Jizhong Han. Explicit correlation learning for generalizable cross-modal deepfake detection, 2024.