SYNC WHEN IT COUNTS: A MODALITY-CONDITIONAL FRAMEWORK FOR ROBUST AUDIO-VISUAL DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Current audio-visual forgery detectors primarily exploit cross-modal alignment, assuming a continuous synchronization between facial movements and the acoustic speech signal. While effective on curated benchmarks, this rigid paradigm catastrophically fails under real-world conditions. When a single modality is missing or degraded, state-of-the-art alignment models (e.g., AVH-Align) collapse to near-random performance. This reveals the vulnerability that enforcing cross-modal correlation effectively blinds the model to the rich, independent forensic signals within each surviving modality. To address this, we propose a modality-conditional, self-supervised framework that completely decouples unimodal representation learning from cross-modal synchronization. Our architecture pairs a contrastive audio-visual alignment module with independent visual and audio Joint-Embedding Predictive Architecture (JEPA) streams trained entirely without manipulation labels. During inference, a deterministic routing gate evaluates the input across two tiers. To prevent modality collapse, it routes inputs with missing modalities directly to the independent JEPA streams. For standard audio-visual inputs, it halts false-positive drift by explicitly dropping natural silences (e.g., authentic pauses or nods) from the synchronization calculation. Our method achieves state-of-the-art in-domain performance on AV-Deepfake1M++ (0.8525 AUC, 0.9434 AP). Crucially, the unimodal JEPA pathways retain robust detection capabilities under severe modality degradation (0.739 audio-only AUC and 0.673 visual-only AUC), whereas prior baselines fail entirely. Furthermore, in zero-shot cross-dataset evaluation on In-DeepFake, our approach achieves 0.764 AUC, outperforming several supervised in-domain models. These findings demonstrate that substituting rigid alignment with conditional, context-aware routing is crucial for resilient deepfake detection in real-world scenarios.

1 INTRODUCTION

The foundation of modern audio-visual deepfake detection rests on a single, compelling intuition: generative models inherently struggle to perfectly synchronize synthesized speech with facial motion (Wang, 2025; Haliassos et al., 2021). State-of-the-art methods have operationalized this insight, achieving formidable performance on curated benchmarks. However, this success is built upon a structurally fragile assumption that real-world video consists of a continuous, perfectly synchronized stream of active speech. In reality, natural human conversation is riddled with semantic silence. Speakers pause, nod, glance, and react, producing perfectly authentic facial motion with no corresponding acoustic signal. Ambient audio and background noise further decouple what is seen from what is heard. Continuous alignment-based detectors, strictly optimized to reward cross-modal correlation, misinterpret these natural asynchronies as evidence of forgery. As illustrated in Figure 1, valid non-speech behaviors are conflated with manipulation artifacts, resulting in severe false-positive drift. Authentic footage is penalized simply because the model was never taught that silence is a natural state. This architectural blind spot results in a second, catastrophic failure mode: *modality collapse*. When a single modality is degraded or entirely missing, state-of-the-art detectors fail completely. For instance, AVH-Align (Smeu et al., 2025) achieves a decent 0.83 AUC (approx.) on AV-Deepfake1M++ (Cai et al., 2025) under ideal conditions yet degrades to a near-random 0.48

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.

AUC (audio-only) and 0.46 AUC (visual-only). SpeechForensics (Liang et al., 2024) fares no better, collapsing to 0.50 AUC in both unimodal scenarios. Since these models depend entirely on cross-modal correlation, they effectively discard the rich, independent forensic traces embedded within the surviving modality. This vulnerability persists even in concurrent 2026 work: X-AVDT (Kim et al., 2026), which relies on speech-driven audio-visual cross-attention, explicitly identifies non-speech and missing-speech conditions as an open weakness and lists unimodal back-off strategies as future work, rather than a solved capability, underscoring that robustness to modality degradation remains an unaddressed gap across current alignment-based approaches.

Thus, to achieve robust, in-the-wild detection, synchronization must be treated as a *conditional* signal: highly discriminative when speech is active, but safely bypassed when it is not. To this end, we propose a modality-conditional self-supervised framework that completely decouples unimodal representation learning from cross-modal synchronization. On one hand, independent visual and audio Joint-Embedding Predictive Architecture (JEPA) streams (Assran et al., 2023) learn robust, standalone representations by predicting masked temporal context, capturing localized spatial jitter and prosodic anomalies without requiring cross-modal pairs. On the other hand, a contrastive synchronization module focuses strictly on cross-modal alignment. Crucially, inference is governed by a deterministic two-tiered routing gate. To prevent modality collapse, completely missing modalities are routed natively to the independent JEPA streams. To halt false-positive drift in standard videos, the gate evaluates per-window signal energy and motion variance, dynamically dropping semantically inactive segments (e.g., pauses or silent reactions) from the synchronization calculation. By ensuring that the model evaluates cross-modal evidence conditionally, we eliminate the penalty for natural silence while retaining robust detection under modality degradation.

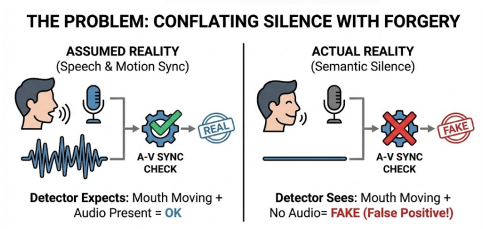


Figure 1: Current audio-visual deepfake detectors assume that facial motion always aligns with speech (Left). This assumption breaks during semantic silence, where authentic non-speech behaviors such as silent CCTV footage occur without audio (Right). Alignment-based models misinterpret this valid absence of synchronization as manipulation, leading to elevated false positives in real-world videos.

Our main contributions are as follows:

- We identify and empirically quantify *modality collapse* and *false-positive drift* as systemic, architectural failure modes in continuous alignment-based deepfake detectors.
- We propose a fully decoupled, modality-conditional self-supervised framework that pairs contrastive audio-visual synchronization with independent JEPA unimodal streams, governed by a deterministic two-tiered routing gate.
- We introduce synchronization standard deviation to capture *temporal sync jitter*, providing a superior test-time aggregation strategy.
- We demonstrate state-of-the-art self-supervised performance, including robust zero-shot generalization on the linguistically diverse InDeepFake benchmark and unprecedented resilience to missing modalities.

2 Related Work

2.1 Audio-Visual Deepfake Detection

Early multimodal forgery detectors treated audio and video as independent streams, fusing CNN or RNN decision scores at the output stage (Zhou et al., 2017). While simple, this design discards the rich temporal interdependencies between modalities. A paradigm shift arrived with SyncNet (Wang, 2025), which demonstrated that learning to match lip motion with speech audio yields powerful authenticity features. This insight was rapidly adapted for detection: LipForensics (Haliassos et al., 2021) leveraged mouth motion embeddings pre-trained on lip-reading to expose unnatural mimicry. SpeechForensics (Liang et al., 2024) extended this by modeling the semantic correlation between

phonemes and visemes as a deviation-based forgery signal. More recently, AVH-Align (Smeu et al., 2025) aligned dense frame-level features from AV-HuBERT (Shi et al., 2022a) to capture fine-grained temporal inconsistencies, achieving state-of-the-art results on benchmarks such as FakeAVCeleb (Khalid et al., 2021) and AV-Deepfake1M (Cai et al., 2024). AVAD (Feng et al., 2023) proposed a self-supervised audio-visual anomaly detection framework, while CAV-MAE (Gong et al., 2023) and X-AVDT (Kim et al., 2026) introduced advanced cross-modal mechanisms for learning joint multimodal representations. Despite strong benchmark numbers, all of these methods enforce synchronization universally, across every frame, regardless of semantic context. This rigid assumption is the central vulnerability our work addresses.

2.2 Robustness and Generalization in Deepfake Detection

High leaderboard scores often mask fragile generalization. Models trained on specific manipulation artifacts frequently fail on unseen synthesis methods (Verdoliva, 2020; Aarthi & Agarwal, 2025). Dataset bias is a recurring problem: Smeu et al. (2025) demonstrated that many benchmarks contain spurious "leading silence" shortcuts that models overfit to, collapsing when silence is redistributed. Similarly, Cozzolino et al. (2023) showed that identity-specific biometric priors, rather than manipulation artifacts, often drive detection decisions, inflating in-domain performance while masking poor generalization. Cross-dataset evaluation has emerged as a necessary stress test (Groh et al., 2022), revealing that most detectors fail dramatically under distribution shift. Our work contributes to this line of inquiry by exposing a previously underquantified failure mode: *modality collapse*, where detectors trained on aligned audio-visual streams degrade to near-random chance when a single modality is absent, corrupted, or naturally silent.

2.3 Semantic Silence and Conditional Synchronization

A critical yet under-explored failure mode is the handling of *semantic silence*, i.e., authentic facial motion (nodding, smiling, reacting) that naturally occurs without corresponding speech. Alignment-based detectors, trained to maximize cross-modal correlation, misinterpret these valid asynchronies as manipulation artifacts, producing elevated false-positive rates in unconstrained video (Cai et al., 2024). This problem is structurally distinct from the "technical silence" shortcuts studied by Smeu et al. (Smeu et al., 2025). This arises not from dataset artifacts, but from the fundamental mismatch between real-world facial dynamics and the always-synchronized training assumption. Conditional or speech-aware architectures have been proposed in the context of audio-visual representation learning (Shi et al., 2022a;b), but have not been explored for deepfake detection. Our gating mechanism directly addresses this gap by evaluating signal energy and motion variance per temporal window, routing semantically inactive segments away from synchronization scoring.

2.4 Self-Supervised Learning for Forgery Detection

The scarcity of labeled manipulation data has driven the adoption of self-supervised learning (SSL) for deepfake detection (Smeu et al., 2025; Liang et al., 2024). Contrastive methods such as SimCLR (Chen et al., 2020), MoCo (He et al., 2020), and SyncLipMAE (Ling et al., 2025) have been adapted for audio-visual alignment, pulling temporally corresponding pairs together in embedding space. However, purely contrastive objectives neglect the internal structure of individual modalities, a limitation that becomes critical when one modality is absent or unreliable. Predictive architectures offer a complementary solution. For example, Masked Autoencoders (MAE) (He et al., 2022) reconstruct masked input regions, enforcing rich structural understanding within a single modality. JEPA (Assran et al., 2023) extends this by predicting masked latent representations rather than pixel-level reconstructions, producing higher-level semantic features. Video-JEPA and related extensions (Assran et al., 2023) have demonstrated strong temporal reasoning from real video alone. In the audio domain, models such as HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), and data2vec (Baevski et al., 2022) learn powerful speech representations through masked prediction, forming the backbone of modern audio forgery detection (Tak et al., 2022). Our approach builds on this foundation but introduces a novel extension: a dual-head JEPA that jointly predicts frame representations and their temporal velocity (discrete difference). This explicitly supervises temporal smoothness, making the model sensitive to the feature-space jitter characteristic of synthesized content. This temporal sensitivity also informs how frame-level scores should be aggregated into a video-level

decision: mean pooling, the most common strategy, dilutes localized artifacts characteristic of partial manipulation (Cai et al., 2024), motivating alternatives such as Multiple Instance Learning pooling or attention-weighted pooling (Zhao et al., 2021). We extend this analysis by introducing *sync score standard deviation* as a novel test-time aggregation statistic.

3 Proposed Methodology for Robust Deepfake Detection

3.1 Problem Formulation and Architectural Overview

We formulate audio-visual deepfake detection as a fully self-supervised anomaly detection task, requiring zero manipulation labels during training. Unlike existing methods that rigidly enforce cross-modal correlation across all inputs, our framework is designed around a fully decoupled, modality-conditional architecture. As illustrated in Figure 2 and the algorithm given in Appendix B, it comprises three independently optimized pathways:

1. A **Contrastive Audio-Visual Synchronization Module**, dedicated to measuring cross-modal temporal alignment using lip-cropped visual features and audio features.
2. A **Dual-Target Visual JEPA Stream**, operating on full-face features to model global facial dynamics and detect localized spatial jitter.
3. An **Audio JEPA Stream**, modeling speech dynamics and prosodic structure purely from acoustic features.

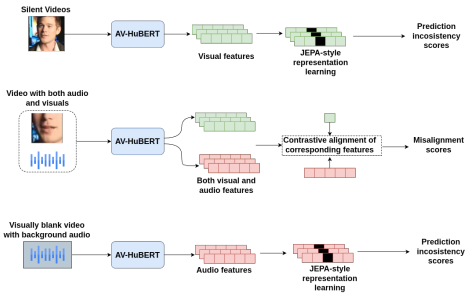


Figure 2: Overview of the proposed conditional audio–visual deepfake detection framework. When both audio and visual modalities are present, the model performs contrastive audio–visual synchronization to produce alignment scores. In the absence of one modality, unimodal JEPA-style predictive modeling is used to compute modality-specific inconsistency scores. An enlarged version of this image is given in the Appendix B.

Crucially, these components share no parameters, are never jointly optimized, and are *not fused* at inference time. Instead, they operate as mutually exclusive evaluation pathways governed by a deterministic routing mechanism, ensuring the system naturally adapts to modality degradation and semantic silence.

3.2 Self-Supervised Feature Extraction

To support our decoupled architecture, we extract two distinct streams of feature representations at 25 frames per second (FPS), yielding 1024-dimensional embeddings per time step.

Synchronization Features (Audio and Lip-Visual). For the synchronization module, we require representations that strongly encode phonetic and articulatory alignment. We select AV-HuBERT (Shi et al., 2022a) as our backbone for this task, as its self-supervised pre-training objective of predicting joint multimodal clusters from masked inputs forces the network to learn a highly correlated latent space where lip kinematics and speech acoustics are inherently coupled. Following established protocols (Smeu et al., 2025), we extract purely unimodal representations from this backbone by masking the complementary stream. We extract the audio sequence $\mathbf{a}_t$ by masking the visual input, and the corresponding visual sequence $\mathbf{v}_t^{\text{lip}}$ by providing cropped lip regions while masking the audio input.

Full-Face Visual Features (For Visual JEPA). While cropped lips are optimal for speech synchronization, they discard critical macro-dynamics (e.g., head pose, eye gaze). To capture these for our Visual JEPA stream, we pass the resized full-face bounding box through AV-HuBERT's initial 3D ResNet-18 spatio-temporal front-end with the audio stream completely masked. This extracts a sequence of rich global spatial visual features $\mathbf{v}_t^{\text{face}}$ that encapsulate the entire facial dynamic context.

3.3 TWO-TIERED INFERENCE ROUTING (MODALITY SELECTION)

To structurally address the "modality collapse" vulnerability, we introduce a parameter-free, two-tiered routing strategy that dictates exactly how a video is scored.

Tier 1: Video-Level Modality Selection. Before processing, the global availability of modalities is evaluated to select the active pathway:

- **Standard Audio-Visual Input:** If a video contains both audio and visual streams, it is routed *exclusively* to the Contrastive Synchronization Module. The JEPA streams are entirely bypassed.
- **Missing Modalities (Unimodal Fallback):** If a modality is completely absent from the outset (e.g., silent CCTV footage), the input is routed natively to the respective active unimodal stream (Visual JEPA or Audio JEPA).

Tier 2: Window-Level Semantic Filtering (Silence Dropping). For standard audio-visual videos evaluated by the synchronization pathway, forcing the model to calculate alignment during natural silence artificially inflates the anomaly score. Therefore, before a temporal window $\mathcal{W}$ contributes to the final synchronization calculation, it is deterministically filtered out if it meets either of the following semantic conditions:

- **Insufficient Audio Energy:** We compute temporal energy from the raw preprocessed acoustic filter banks. If the average filterbank energy across the window falls below ϵ_a, it is classified as a speech pause.
- **Insufficient Visual Activity:** We compute the temporal variance of the ℓ_2 norms of the visual embeddings ($\mathrm{Var}_{t\in\mathcal{W}}(\|\mathbf{v}_t^{\text{lip}}\|_2) < \epsilon_v$). While this measure features magnitude variance rather than explicit coordinate displacement (preventing constant-velocity camera pans from erroneously evaluating to zero), fluctuations in this deep latent space strongly correlate with active articulation.

To ensure robustness across diverse recording conditions, the thresholds ϵ_a and ϵ_v are not arbitrary static hyperparameters; they are calibrated dynamically at the 5th percentile of the respective energy and variance distributions computed over a small, unlabelled validation set of authentic videos. Windows triggering these conditions are simply *omitted* from the synchronization pathway. The final video-level synchronization score is aggregated strictly from windows containing valid, active speech. Because ϵ_a and ϵ_v are calibrated on a validation split rather than fixed a priori, we additionally verify that this calibration transfers across domains: thresholds calibrated on AV-Deepfake1M++ and applied unchanged to InDeepFake retain within one AUC point of natively recalibrated thresholds while filtering substantially more low-information windows before the transformer stage (Appendix F).

3.4 CONTRASTIVE AUDIO-VISUAL SYNCHRONIZATION (THE PRIMARY PATHWAY)

Training Objective. The synchronization module is trained exclusively on authentic video. For each valid time step t, we define a bidirectional window of size $2\tau + 1$ containing a central visual feature $\mathbf{v}_t^{\text{lip}}$ and corresponding audio features $\{\mathbf{a}_{t+k}\}_{k=-\tau}^{\tau}$. Modalities are mapped into a shared ℓ_2-normalized embedding space ($\mathbf{z}_t^v$, $\mathbf{z}_{t+k}^a$) via independent two-layer MLP encoders.

Frame-level alignment is quantified via scaled cosine similarity $s_{t,k} = (\mathbf{z}_t^v \cdot \mathbf{z}_{t+k}^a)/\mathcal{T}$. The network is optimized using an InfoNCE contrastive loss, treating the temporally aligned audio frame ($k = 0$) as the positive pair and all shifted frames within the window as negatives:

$$\mathcal{L}_{\text{sync}} = -\mathbb{E}_t \left[\log \frac{\exp(s_{t,0})}{\sum_{k=-\tau}^{\tau} \exp(s_{t,k})} \right] \tag{1}$$

Test-Time Scoring and Aggregation. For an evaluated video, each active speech window yields an aligned sync score $s_t = s_{t,0}$. After median-filtering (kernel size 5) to suppress isolated noise, we aggregate the sequence $\{s_t\}_{t=1}^{T}$ into a final video decision using Standard Deviation ($S_{\text{std}} = \sigma(\{s_t\})$). The standard deviation directly captures *sync jitter*; the temporal instability and variance created when

deepfake frame-blending algorithms splice segments together. This choice is grounded in the underlying failure mode rather than surface heuristics: a local-variance analysis shows synchronization instability peaks *within* manipulated content itself (mean window STD $= 2.230$) rather than at splice boundaries (mean STD $= 1.879$, Mann-Whitney $p < 0.001$), and a partial Spearman correlation controlling for active-window count finds this instability essentially independent of edit quantity ($r = 0.0010$, $p = 0.93$; Appendix G). Together these indicate S_{std} captures a systemic failure of generative models to hold temporal latents stable, rather than an artifact of stitching that could be smoothed away in post-processing.

3.5 Unimodal Representation Learning (The Fallback Pathways)

When a video lacks a modality (Tier 1 routing), it is evaluated by our JEPA (Assran et al., 2023) streams, trained strictly as one class on real data.

Visual JEPA (Dual-Target). Operating on full-face features to capture global dynamics, the visual stream analyzes an ℓ_2-normalized window $\mathbf{V} = \{\mathbf{v}_{t-\tau}^{\text{face}}, \ldots, \mathbf{v}_{t+\tau}^{\text{face}}\}$. The center frame is replaced by a learned mask token and processed via a 6-layer Transformer encoder. Unlike standard JEPAs requiring momentum teachers, our targets are derived directly from static unmasked features.

Two independent heads predict the masked information: the *Representation Head* predicts the unmasked feature ($\hat{\mathbf{z}}_v = \mathbf{v}_t^{\text{face}} / \|\mathbf{v}_t^{\text{face}}\|_2$), and the *Gradient Head* predicts the temporal velocity ($\hat{\boldsymbol{\delta}}_t = (\mathbf{v}_t^{\text{face}} - \mathbf{v}_{t-1}^{\text{face}}) / \|\mathbf{v}_t^{\text{face}} - \mathbf{v}_{t-1}^{\text{face}}\|_2$).

The network minimizes both prediction errors:

$$\mathcal{L}_{\text{JEPA}}^{v} = \left\| f_{\text{repr}}(\tilde{\mathbf{V}}) - \hat{\mathbf{z}}_v \right\|_2^2 + \lambda \left\| f_{\text{grad}}(\tilde{\mathbf{V}}) - \hat{\boldsymbol{\delta}}_t \right\|_2^2 \tag{2}$$

The gradient head explicitly supervises the rate of change in the feature space, identifying the abrupt temporal jumps characteristic of generative spatial manipulation. Frame-level errors (e_t) are aggregated at the 95th percentile and calibrated as upper-tail positive z-deviations relative to the authentic distribution.

Audio JEPA. Following an identical predictive setup, the audio stream masks and predicts the center frame of an audio window to learn prosodic structure and pause distributions, targeting the acoustic feature $\hat{\mathbf{z}}_a = \mathbf{a}_t / \|\mathbf{a}_t\|_2$:

$$\mathcal{L}_{\text{JEPA}}^{a} = \left\| f_{\text{repr}}(\tilde{\mathbf{A}}) - \hat{\mathbf{z}}_a \right\|_2^2 \tag{3}$$

Test-time scoring mirrors the visual stream via calibrated 95th-percentile aggregation.

3.6 Experimental Setup

All architectural components are optimized in a strictly self-supervised manner on AV-Deepfake1M++ (Cai et al., 2025). Training utilizes 45,000 authentic videos, with 5,000 reserved for validation and early stopping. Final evaluation is conducted on 8,930 held-out videos containing both real and manipulated samples. All models are optimized using Adam (lr $= 1 \times 10^{-5}$). The contrastive synchronization module is configured with $\tau = 8$ (lip-cropped), while both JEPA streams operate at $\tau = 15$. Performance is reported using Area Under the Receiver Operating Characteristic Curve (AUC) and Average Precision (AP).

4 Results and Analysis

4.1 In-Domain Performance

As our framework employs deterministic routing rather than continuous multi-stream fusion, video-level scoring is architecturally straightforward. A video's final anomaly score is simply the score from its natively selected pathway (e.g., the standard deviation of sync scores for standard audio-visual inputs, or the 95th-percentile JEPA error for unimodal fallback cases). Table 1 reports the performance

Table 1: Performance of self-supervised audio-visual deepfake detectors on AV-Deepfake1M++. All models are trained exclusively on real video.

Method	Modality	Supervision	AUC ↑	AP ↑
AVAD (Feng et al., 2023)	AV	Self-sup.	0.5397	0.7597
SpeechForensics (Liang et al., 2024)	AV	Self-sup.	0.6627	0.8219
AVH-Align (Smeu et al., 2025)	AV	Self-sup.	0.8274	0.9295
Ours	AV	Self-sup.	**0.8525**	**0.9434**

Table 2: Unimodal robustness evaluation. All three conditions are evaluated on the identical held-out set of 8,930 AV-Deepfake1M++ videos; missing modalities are obtained by masking the corresponding input stream at inference time, so visual/audio content and manipulation methods are held fixed across conditions, and modality collapse is isolated as the sole variable. Missing modalities in baselines are mocked with a neutral mean tensor.

Method	AV AUC	Audio-Only AUC	Visual-Only AUC
SpeechForensics (Liang et al., 2024)	0.6600	0.5000	0.5000
AVH-Align (Smeu et al., 2025)	0.8274	0.4800	0.4600
Ours	**0.8525**	**0.7392**	**0.6727**

of state-of-the-art self-supervised detectors on AV-Deepfake1M++. Our method achieves **0.8525 AUC** (95% CI: [0.844, 0.861]) and **0.9434 AP** (95% CI: [0.938, 0.948]). Crucially, the lower bound of our 95% confidence interval strictly supersedes the absolute performance of the previous state-of-the-art, AVH-Align, definitively proving that our conditional gating and variance-based aggregation provide statistically significant discriminative gains.

4.2 Modality Collapse and Unimodal Robustness

Table 2 presents the most critical empirical finding of this paper: the phenomenon of *modality collapse*. While AVH-Align achieves 0.8274 AUC under ideal conditions, it completely collapses to near-random chance (0.48 audio-only and 0.46 visual-only AUC) when a single modality is degraded. Since these baseline architectures strictly enforce cross-modal correlation, they are structurally incapable of reasoning about a modality in isolation. Our decoupled architecture solves this fundamental vulnerability. By routing degraded streams exclusively to the independent JEPA modules, our method retains a robust **0.739 AUC audio-only** and **0.673 AUC visual-only**. This ablation proves that our JEPA streams successfully extract rich, independent forensic traces (such as unnatural temporal feature velocity) that continuous alignment models inherently discard. A bootstrap analysis (1,000 resamples) confirms these gains are stable: our AV, audio-only, and visual-only AUCs of 0.8525, 0.7392, and 0.6727 carry 95% confidence intervals of [0.8439, 0.8605], [0.7255, 0.7516], and [0.6664, 0.6922] respectively, with none overlapping the corresponding baseline scores.

Why routing instead of fusion: We evaluate an unconditional Max-Fusion baseline against our decoupled design; fusion consistently degrades detection across all forgery categories, dropping overall AUC from 0.8525 to 0.8046 (per-subtype breakdown in Appendix E.

4.3 Bypassing Synchronization Penalties During Semantic Silence

Table 3 evaluates performance on the silent video subset of AV-Deepfake1M++. Continuous alignment models actively penalize authentic non-speech behaviors hovering near 0.50 AUC. Generic visual foundation models (DINOv2 (Oquab et al., 2024), Video-MAE (Tong et al., 2022)) evaluated using identical one-class anomaly scoring protocols (upper-tail Z-scores calibrated on real data) for fair comparison fare only marginally better (For exhaustive details regarding the exact evaluation protocols used for these baselines, please refer to Appendix A.). While absolute AUC is inherently bounded on this subset due to the complete absence of acoustic forensic cues, the primary objective is halting false-positive drift. Our deterministic gating ensures that valid silent motion is explicitly bypassed

Table 3: Detection performance on the strictly silent video subset of AV-Deepfake1M++.

METHOD	MODALITY	AUC ↑	AP ↑
VIDEO-MAE (TONG ET AL., 2022)	V	0.5290	0.5389
FACE-MAE (WANG ET AL., 2022)	V	0.5503	0.5357
AVAD (FENG ET AL., 2023)	AV	0.5000	0.5032
DINOV2 (OQUAB ET AL., 2024)	V	0.5234	0.5298
AU-GUIDED (ANAND ET AL., 2025)	V	0.5018	0.4861
FSDDETECTOR (NGUYEN ET AL., 2025)	V	0.5140	0.4854
CAV-MAE SYNC (ARAUJO ET AL., 2025)	AV	0.5125	0.5082
AVH-ALIGN (SMEU ET AL., 2025)	AV	0.5462	0.5324
OURS	V	**0.5651**	**0.5877**

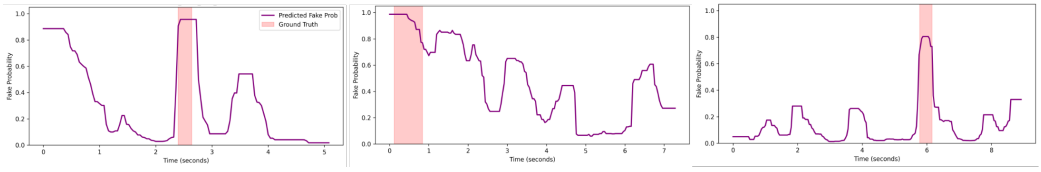


Figure 3: Temporal localization of manipulation evidence using synchronization-based scoring. The predicted fake probability exhibits sharp, localized peaks aligned with manipulated regions (shaded), while remaining low during real segments.

by the synchronization module and evaluated purely by the visual JEPA. Consequently, our method avoids the cross-modal penalty and establishes a consistent performance floor (**0.5651 AUC**, 95% CI: [0.5088, 0.6752]). The wider interval here reflects the smaller size of the silent-video subset relative to the full test set.

4.4 ABLATION: THE MECHANICS OF AGGREGATION AND CONTEXT LENGTH

Standard-deviation aggregation dominates mean and low-percentile pooling across all tested context lengths, and shorter windows outperform longer ones (τ=8: 0.8525 AUC vs. τ=24: 0.8239 AUC), confirming that deepfake synchronization artifacts operate at a fine temporal granularity that wider windows dilute (full ablation grid in Appendix I).

4.5 ZERO-SHOT GENERALIZATION: OVERCOMING SYNTHESIS DISTRIBUTION SHIFTS

Table 4 evaluates zero-shot generalization on InDeepFake (Das et al., 2025), a multilingual benchmark spanning seven Indian languages with completely unseen manipulation engines. Our method achieves an exceptional **0.764 AUC** zero-shot, outperforming all supervised in-domain baselines and outperforming all self-supervised baselines, with a bootstrap 95% CI of [0.6848, 0.8540] on this result. In fact, continuous alignment methods (AVH-Align, AVAD) also fail to generalize (0.56 and 0.58 AUC) to InDeepFake. For all zero-shot evaluations, the anomaly score polarity is strictly determined by the model's pre-training objective: distance and variance metrics are treated as positive indicators of manipulation, while similarity-based scores are negated prior to AUC computation. This phenomenon of supervised dataset bias is not limited to audio-visual models; we present a dedicated analysis of catastrophic cross-domain collapse in supervised audio-only detectors in Appendix C.

4.6 TEMPORAL LOCALIZATION OF AUDIO-VISUAL INCONSISTENCIES

Beyond video-level detection accuracy, effective deepfake detectors should localize when manipulation artifacts occur. Real-world deepfakes rarely exhibit uniform degradation across time. Instead, artifacts emerge intermittently due to imperfect lip generation, temporal drift, or inconsistent audio-visual alignment. A detector that distributes suspicion uniformly across a video obscures this structure and reduces interpretability. Figure 3 illustrates the temporal behavior of our synchronization-based detector on representative examples. The predicted fake probability remains low during well-aligned

Table 4: Zero-shot cross-dataset evaluation on InDeepFake (Das et al., 2025). Our self-supervised method outperforms both supervised in-domain models and recent zero-shot baselines across modalities.

Method	Modality	Supervision	Training Data	AUC ↑
Capsule	V	Supervised	InDeepFake	0.5700
MesoInception4	V	Supervised	InDeepFake	0.6400
Xception-comp	V	Supervised	InDeepFake	0.7100
Meso4	V	Supervised	InDeepFake	0.7200
AU-Guided (Anand et al., 2025)	V	Supervised	Cross-Domain	0.6560
FSDDetector (Nguyen et al., 2025)	V	Supervised	Cross-Domain	0.6701
AVAD (Feng et al., 2023)	AV	Self-sup.	LRS2, LRS3	0.5800
CAV-MAE Sync (Araujo et al., 2025)	AV	Self-sup.	AudioSet	0.5280
AVH-Align (Smeu et al., 2025)	AV	Self-sup.	AV-Deepfake1M	0.5600
HAVIC (Peng et al., 2026)	AV	Self-sup.	LRS2, FakeAVCeleb	0.5137
(Anshul et al., 2026)	AV	Self-sup.	VoxCeleb2, FakeAVCeleb	0.5154
Ours	AV	Self-sup.	AV-Deepfake1M++	**0.7640**

segments and rises sharply during short temporal intervals corresponding to manipulated regions. Importantly, these peaks are temporally localized rather than persistent, reflecting the sparse nature of audio-visual inconsistencies in realistic deepfakes. Outside these intervals, confidence decays rapidly, indicating that the detector does not over-penalize surrounding authentic content. Qualitative support for isolating localized inconsistencies is further evidenced by our embedding space analysis in Appendix D. Beyond localization accuracy, our primary pathway also meets real-time deployment constraints (345.6 ms per 5-second clip), matching AVH-Align's latency, while the worst-case unimodal fallback remains bounded under 2 s even for 15-second clips (Appendix H).

4.7 Limitations and Future Work

While our conditional synchronization framework significantly improves robustness against semantic silence and modality degradation, some challenges remain open. In real-world environments, audio or visual streams may exhibit partial absence or severe degradation due to prolonged occlusions, sensor failures, or extreme compression artifacts. While our deterministic routing gate natively handles complete modality absence from the outset, dynamically shifting between active synchronization and unimodal fallback during high-frequency, intermittent degradation within a single video sequence requires further exploration. Relatedly, occlusion and non-frontal poses degrade the visual synchronization signal in a manner similar to modality absence, but our current gate only handles complete, not partial or intermittent, signal loss. We view extending the conditional-routing principle from facial/speech synchrony to more general action-onset/sound-onset correlation in unconstrained video as a promising direction for generalizing beyond talking-head deepfakes. Moreover, natural temporal lag and micro-asynchronies can arise legitimately from recording latency, editing pipelines, or transmission delays. While our standard deviation aggregation metric is resilient to global offsets, it may still flag videos with highly volatile, yet authentic, transmission jitter. Extending the framework to explicitly reason about temporal uncertainty and dynamic modality confidence during inference is a promising direction for future research.

5 Conclusion and Broader Impact

This work identifies a structural limitation in alignment-based audio-visual deepfake detectors. They tend to misread natural non-speech behavior and missing modalities as manipulation. We address this by treating cross-modal synchronization as a conditional signal rather than a strict constraint. Our framework pairs this signal with independent unimodal JEPA streams via a deterministic routing gate. This prevents false-positive drift during natural silence and maintains robust detection under severe modality degradation. We validated these improvements across AV-Deepfake1M++ and zero-shot evaluations on InDeepFake. Beyond deepfakes, this context-aware approach benefits real-world self-supervised multimodal systems, paving the way for more reliable multimedia forensics.

Reproducibility Statement

Architecture, training objectives, and the two-tiered inference routing are fully specified in Section 3, with Algorithm 1 and an enlarged schematic in Appendix B. Dataset splits, optimizer settings, and window hyperparameters are given in Section 3.6; threshold calibration and its cross-domain robustness are detailed in Section 3.3 and Appendix F. All reported results for our method include 95% bootstrap confidence intervals (1,000 resamples). Baseline evaluation protocols, including zero-shot inference procedures for models without native missing-modality handling, are given in Appendix A. Code, evaluation scripts, and pre-trained weights will be released publicly upon publication.

AI Use Statement

In this work, we used generative AI tools to refine the phrasing, improve readability, and format the LaTeX source code. We also used an AI image-generation tool to produce illustrative elements within Figure 1. All core ideas, experimental designs, and data analyses are our original work. We carefully reviewed, edited, and approved all AI-assisted text and imagery, and we take full responsibility for the final content of this paper.

References

S Aarthi and Akshay Agarwal. Unmasking the audio illusion: A survey on spoofing and deepfake detection. In *2025 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–11. IEEE, 2025.

Tharun Anand, Siva Sankar, and Pravin Nair. Detecting localized deepfake manipulations using action unit-guided video representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4341–4351, 2025.

Ashutosh Anshul, Eng Siong Chng, and Deepu Rajan. Av representation learning via audio shift prediction for multimodal deepfake detection and temporal localization. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 2553–2563. IEEE, 2026.

Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R Glass, and Hilde Kuehne. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18794–18803, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8519–8529, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pp. 1298–1312. PMLR, 2022.

Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 7414–7423, 2024.

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. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13686–13691, 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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

Arnab Kumar Das, Aritra Bose, Priya Manohar, Anurag Dutta, Ruchira Naskar, and Rajat Subhra Chakraborty. Indeepfake: A novel multimodal multilingual indian deepfake video dataset. *Pattern Recognition Letters*, 197:16–23, 2025. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2025.07.002. URL https://www.sciencedirect.com/science/article/pii/S0167865525002545.

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

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. In *International Conference on Learning Representations*, 2023.

Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119, 2022.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

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.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In *Advances in Neural Information Processing Systems (NeurIPS Datasets and Benchmarks Track)*, 2021. URL https://openreview.net/forum?id=TAXFsg6ZaOl.

Taewoo Kim, Guisik Kim, Choongsang Cho, and Young Han Lee. Naturalness-aware curriculum learning with dynamic temperature for speech deepfake detection. In *Interspeech*, pp. 5318–5322, 2025.

Youngseo Kim, Kwan Yun, Seokhyeon Hong, Sihun Cha, Colette Suhjung Koo, and Junyong Noh. X-avdt: Audio-visual cross-attention for robust deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4403–4414, June 2026.

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. *Advances in Neural Information Processing Systems*, 37:86124–86144, 2024.

Zeyu Ling, Xiaodong Gu, Jiangnan Tang, and Changqing Zou. Synclipmae: Contrastive masked pretraining for audio-visual talking-face representation. *arXiv preprint arXiv:2510.10069*, 2025.

Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of ai-generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3040–3050, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Jielun Peng, Yabin Wang, Yaqi Li, Long Kong, and Xiaopeng Hong. Leave no stone unturned: Uncovering holistic audio-visual intrinsic coherence for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6655–6666, June 2026.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022a.

Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. Robust self-supervised audio-visual speech recognition. In *Interspeech*, 2022b.

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

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop (Odyssey)*, 2022.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Jiankang Deng, Xinchao Wang, Hakan Bilen, and Yang You. Facemae: Privacy-preserving face recognition via masked autoencoders. *arXiv preprint arXiv:2205.11090*, 2022.

Meng Wang. Deepfake detection: A multimodal survey. In *ITM Web of Conferences*, volume 78. EDP Sciences, 2025.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194, 2021.

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

A Extended Baseline Evaluation Protocols

To ensure rigorous benchmarking and definitively address the systemic generalization failures of deepfake detectors across distribution shifts, we evaluated multiple recent state-of-the-art baselines spanning visual and audio-visual modalities. To ensure complete transparency, the pre-training paradigms and exact zero-shot inference protocols for each baseline are detailed below.

To evaluate recent supervised spatial and spatio-temporal baselines, we tested FSDDetector (Nguyen et al., 2025), a zero-shot AI-image detector, alongside an Action Unit-Guided architecture (Anand et al., 2025). Given their distinct architectural requirements, we tailored the zero-shot inference

protocols accordingly. For the spatial FSDDetector, we uniformly sampled 50 frames across the temporal axis of each video, computed independent spatial anomaly z-scores for each frame, and averaged them to generate a video-level score calibrated via a sigmoid function. Conversely, the spatio-temporal Action Unit model evaluated videos natively as 3D spatio-temporal tensors using its official pre-trained weights, directly extracting the spatial-temporal softmax logit corresponding to the "fake" class as the final video score.

Visual Foundation Models (DINOv2, Video-MAE, Face-MAE), *Protocol:* Because these generic representation models lack manipulation-specific classification heads, we instituted an unsupervised temporal smoothness heuristic. We extracted dense frame-level feature embeddings and utilized the inverse frame-to-frame cosine similarity as a proxy for temporal anomaly. Higher scores indicate unnatural structural velocity (jitter) typical of deepfake frame blending.

As detailed in Section 4, despite these exhaustive protocols, supervised spatial detectors and standard foundation models hover near random chance ($\sim$0.50 AUC) on the AV-Deepfake1M++ silent subset. This proves that supervised models severely overfit to spatial artifacts from their training domains, confirming that our Visual JEPA's explicit temporal gradient supervision is architecturally necessary for detecting generative frame-blending.

We benchmarked against CAV-MAE Sync, a state-of-the-art contrastive audio-visual masked autoencoder pre-trained extensively on AudioSet. *Protocol:* To evaluate fine-grained synchronization natively, we implement a strict 16-chunk evaluation sequence. For each video, we uniformly sample 16 center-cropped visual frames. To guarantee perfect modality alignment, we extract 16 corresponding 1-second audio chunks (mapped to 128-bin Kaldi filterbanks) centered exactly on the visual frame timestamps. The model mapped these synchronized pairs into high-dimensional tokens. We compute the cross-modal cosine similarity at each of the 16 timesteps and aggregate the 5th percentile score to isolate the worst-case synchronization collapse. The baseline's poor generalization (0.5013 AUC on AV-Deepfake1M++ and 0.5280 AUC on InDeepFake) underscores that generic audio-visual pre-training does not reliably transfer to generative deepfake asynchrony without deterministic routing.

Computational Resources. All experiments, including model training and baseline evaluations, were conducted on an internal cluster equipped with 2$\times$ NVIDIA H200 NVL GPUs with 141GB of VRAM.

B Extended Architectural Details

While Figure 2 in the main text provides a high-level overview of our modality-conditional framework, we provide an enlarged, high-resolution schematic in Figure 4 for clearer visualization along with algorithm 1 for better understanding.

C Cross-Domain Robustness: The Supervised Generalization Gap

A recurring vulnerability in deepfake detection is catastrophic distribution shift: supervised models perfectly memorize their training domains but fail entirely on in-the-wild synthesis. To explicitly quantify this bias, we evaluated a state-of-the-art supervised audio detector, the XLS-R Conformer with Naturalness-Aware Curriculum Learning (Kim et al., 2025), against our self-supervised Audio JEPA.

The XLS-R Conformer was fully supervised on the ASVspoof 2019 LA dataset (2,580 bona fide and 22,800 spoofed utterances). In contrast, our Audio JEPA was trained strictly one-class using only the 2,580 bona fide utterances of the ASVspoof 2019 LA dataset, actively handicapping it from learning known spoofing distributions.

As shown in Table 5, the supervised baseline predictably maximizes performance on in-domain ASVspoof datasets (> 0.99 AUC). However, when exposed to the modern, highly heterogeneous audio tracks of AV-Deepfake1M++, its performance catastrophically collapses to 0.6444 AUC. Notably, our self-supervised Audio JEPA, when restricted to the identical ASVspoof training domain, experiences a similar collapse (0.5297 AUC).

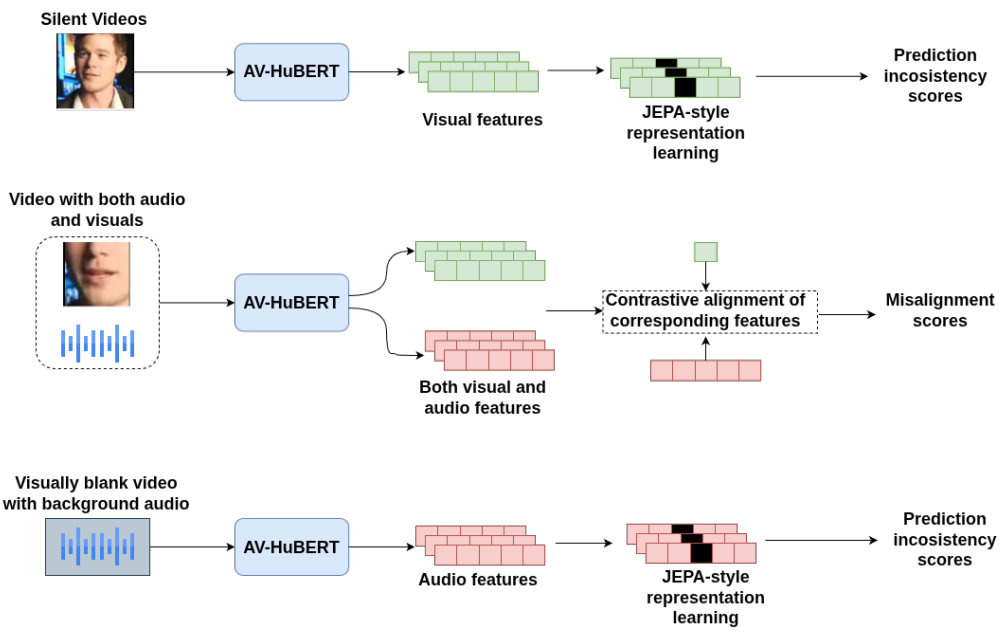


Figure 4: Overview of the proposed conditional audio–visual deepfake detection framework. When both audio and visual modalities are present, the model performs contrastive audio–visual synchronization to produce alignment scores. In the absence of one modality, unimodal JEPA-style predictive modeling is used to compute modality-specific inconsistency scores.

Table 5: Cross-dataset evaluation illustrating the profound domain gap between legacy acoustic spoofing datasets and modern audio-visual deepfakes. Both supervised and self-supervised models trained on ASVspoof 2019 LA suffer catastrophic performance drops when evaluated on the LLM-driven audio of AV-Deepfake1M++, highlighting the necessity of domain-specific representation learning.

	XLS-R Conformer (Supervised)		**Ours (Audio JEPA)**	
Test Dataset	AUC ↑	AP ↑	AUC ↑	AP ↑
ASVspoof 2019 LA (Complete)	1.0000	1.0000	0.7117	0.9548
ASVspoof 2021 DF (50k Subset)	0.9964	0.9999	0.7064	0.9838
ASVspoof 2021 DF (1k Subset)	0.9963	0.9999	0.7131	0.9843
ASVspoof 2021 LA (1k Subset)	0.9829	0.9649	0.5616	0.5325
ASVspoof 2021 PA (1k Subset)	0.6977	0.9201	0.5856	0.8876
AV-Deepfake1M++	**0.6444**	**0.6124**	**0.5297**	**0.5200**

This mutual degradation is a critical finding: it definitively proves that the acoustic artifacts present in legacy voice-spoofing datasets (ASVspoof) are fundamentally disconnected from the generative artifacts found in modern audio-visual synthesis pipelines. This domain mismatch explicitly justifies our architectural decision to train our primary models natively on AV-Deepfake1M++ (which recovers a robust 0.739 AUC for our Audio-Only JEPA, as reported in Table 2), rather than relying on legacy acoustic pre-training.

Algorithm 1 Two-Tiered Conditional Inference for Deepfake Detection

Require: Video input with potential Visual ($\mathbf{V}$) and/or Audio ($\mathbf{A}$) streams.
Require: Calibrated activity thresholds ϵ_a (audio) and ϵ_v (visual).
Require: Pre-trained components: Contrastive Sync module (f_{sync}), Visual JEPA ($f_{\text{v-jepa}}$), Audio JEPA ($f_{\text{a-jepa}}$).
Ensure: Video-level anomaly score S_{final}.

% Tier 1: Video-Level Modality Selection
if $\mathbf{V}$ is available **and** $\mathbf{A}$ is available **then**
  $\mathcal{S} \leftarrow \emptyset$ {Initialize active speech sync scores array}
  Extract sequence features $\mathbf{v}_t^{\text{lip}}$ and $\mathbf{a}_t$ via AV-HuBERT
  for each temporal window $\mathcal{W}_t$ of size $2\tau + 1$ **do**
    % Tier 2: Window-Level Semantic Filtering
    Compute average audio energy: $E_a = \text{mean}_{k \in \mathcal{W}_t}(\text{Energy}(\mathbf{a}_k))$
    Compute visual kinematic variance: $\sigma_v = \text{Var}_{k \in \mathcal{W}_t}(\|\mathbf{v}_k^{\text{lip}}\|_2)$
    if $E_a < \epsilon_a$ **or** $\sigma_v < \epsilon_v$ **then**
      Continue {Skip semantic silence / inactive windows}
    else
      Compute contrastive alignment: $s_t = f_{\text{sync}}(\mathbf{v}_t^{\text{lip}}, \{\mathbf{a}_{t+k}\}_{k=-\tau}^{\tau})$
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{s_t\}$
    end if
  end for
  $\mathcal{S}_{\text{smooth}} \leftarrow \text{MedianFilter}(\mathcal{S}, \text{kernel_size} = 5)$
  $S_{\text{final}} \leftarrow \text{StandardDeviation}(\mathcal{S}_{\text{smooth}})$ {Measures Temporal Sync Jitter}
else if $\mathbf{V}$ is available **and** $\mathbf{A}$ is absent **then**
  {Visual-Only Fallback}
  Extract full-face global features $\mathbf{v}_t^{\text{face}}$
  $\mathcal{E}_v \leftarrow \emptyset$
  for each temporal window $\mathcal{W}_t$ **do**
    Compute dual-target error: $e_t = f_{\text{v-jepa}}(\mathcal{W}_t)$
    $\mathcal{E}_v \leftarrow \mathcal{E}_v \cup \{e_t\}$
  end for
  $S_{\text{final}} \leftarrow 95^{\text{th}}$ Percentile$(\mathcal{E}_v)$ {Measures localized spatial jitter}
else if $\mathbf{A}$ is available **and** $\mathbf{V}$ is absent **then**
  {Audio-Only Fallback}
  Extract acoustic features $\mathbf{a}_t$
  $\mathcal{E}_a \leftarrow \emptyset$
  for each temporal window $\mathcal{W}_t$ **do**
    Compute acoustic error: $e_t = f_{\text{a-jepa}}(\mathcal{W}_t)$
    $\mathcal{E}_a \leftarrow \mathcal{E}_a \cup \{e_t\}$
  end for
  $S_{\text{final}} \leftarrow 95^{\text{th}}$ Percentile$(\mathcal{E}_a)$ {Measures prosodic anomaly}
end if
return S_{final} =0

D Embedding Space Analysis

To gain insight into the structure of the learned audio-visual representations, we visualize video-level embeddings using t-SNE as illustrated in Figure 5. Each point corresponds to a video representation derived from the synchronization module's temporal scoring. Specifically, because videos vary in temporal length, we construct a fixed-dimensional feature vector for each video by capturing its average cross-modal similarity distribution. For every active temporal window, the model computes a $(2\tau + 1)$-dimensional vector of scaled cosine similarities between the central visual feature and the surrounding audio features. We temporally average these frame-level similarity curves across the entire video sequence to form a single feature vector of size $2\tau + 1$ (e.g., 17 dimensions for $\tau = 8$). This vector concisely captures the structural alignment distribution of the video and is directly projected into the 2D space via t-SNE, with colors indicating real and fake samples. While t-SNE

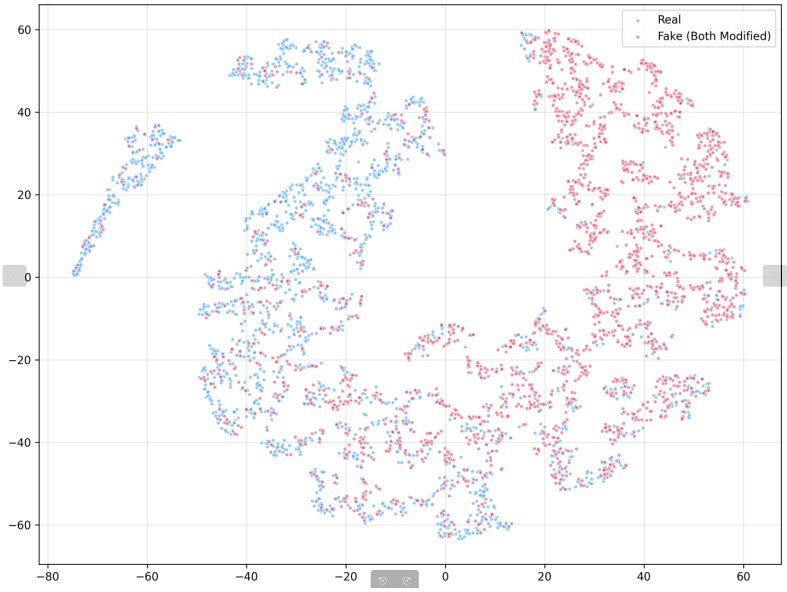


Figure 5: t-SNE visualization of audio-visual synchronization embeddings for real and fake videos. Partial overlap highlights the limitations of unconditional synchronization cues and motivates aggregation strategies that emphasize localized inconsistencies (temporal jitter) rather than global separability. "Fake (Both Modified)" refers to deepfake videos with both audio and visuals manipulated.

does not preserve global distances, it provides a highly useful qualitative view of local neighborhood structure and class overlap.

The visualization reveals several notable patterns. First, real videos form relatively coherent and continuous manifolds, suggesting consistent synchronization dynamics across diverse identities and recording conditions. Fake videos, in contrast, exhibit significant dispersion and fragment into multiple sub-clusters. This behavior explicitly visualizes the heterogeneous nature of manipulation artifacts, which arise from differing generation pipelines, temporal inconsistencies, and degrees of audio-visual mismatch. Second, partial overlap between real and fake samples persists in several regions of the embedding space. This overlap perfectly aligns with our empirical findings that absolute synchronization alone is an imperfect proxy for authenticity. In particular, videos containing natural semantic silence or weakly expressive facial motion occupy regions that overlap with manipulated samples, underscoring the structural limitations of unconditional alignment-based reasoning. Importantly, the separation between real and fake samples improves drastically along directions corresponding to localized synchronization failures, rather than broad, global shifts in the representation space.

E Routing versus Fusion Ablation

To empirically test whether combining all available signals captures complementary forensic evidence, we evaluated an unconditional Max Fusion strategy. This baseline simultaneously fuses the cross modal synchronization signal with both unimodal JEPA scores. We compare this against our decoupled Pure Sync baseline across all 8930 videos in our held-out test set, deliberately segmenting the evaluation by specific manipulation types to expose exactly where fusion helps or hurts.

As Table 6 demonstrates, the empirical results unequivocally validate our decoupled design. Rather than recovering missing evidence, unconditional fusion actively dilutes the highly discriminative synchronization signal across every single forgery category.

Table 6: Pure Sync versus Max-Fusion AUC across forgery subtypes.

Forgery Subtype	Pure Sync AUC	Max-Fusion AUC	95% CI of Difference
Visual Only Forged	0.7774	0.7150	$[-0.075, -0.050]$
Audio Only Forged	0.9060	0.8610	$[-0.054, -0.036]$
Both Forged	0.8731	0.8364	$[-0.047, -0.026]$
Overall Dataset	0.8525	0.8046	$[-0.057, -0.039]$

Since the confidence intervals for the performance difference are strictly below zero across the board, we have definitive proof that forcing the system to unconditionally fuse intra-modal and cross-modal streams actively harms detection. Our conditional routing is therefore the optimal architecture: it relies on unimodal JEPAs only when the primary sync signal structurally collapses, preserving the superior performance on intact videos while safely rescuing missing modality cases.

F Cross-Domain Robustness of the Tier-2 Gate Thresholds

The energy and variance thresholds ϵ_a and ϵ_v that govern window-level semantic filtering are calibrated at the 5th percentile of the activity distribution on a small, unlabelled validation set of authentic videos, rather than fit as free hyperparameters. Because this calibration set could differ from the deployment domain in practice, we evaluate how sensitive detection performance and computational efficiency are to transferring thresholds calibrated on one dataset to another, unseen one.

Table 7 compares two protocols on InDeepFake: *Native*, where ϵ_a and ϵ_v are recalibrated directly on InDeepFake's own validation split, versus *Transfer*, where the thresholds calibrated on AV-Deepfake1M++ are applied unchanged. We report both AUC and the Window Drop Rate—the fraction of temporal windows filtered out by the Tier-2 gate before reaching the synchronization module.

Table 7: Cross-domain threshold ablation. Native calibration requires target-domain validation data; Transfer applies thresholds calibrated on AV-Deepfake1M++ directly to InDeepFake with no additional data or tuning.

Evaluation Protocol	AUC	Window Drop Rate
AV-Deepfake1M++ Test (Native)	0.8525	2.2%
AV-Deepfake1M++ Silent (Native)	0.0000	100.0%
InDeepFake (Native)	0.771	1.4%
InDeepFake (Transfer)	0.764	19.2%

Native calibration on InDeepFake yields only a marginal AUC improvement over transfer (0.771 vs. 0.764, $\Delta = 0.007$), at the cost of requiring target-domain validation data and additional calibration overhead. The Transfer protocol, by contrast, drops a substantially larger fraction of windows (19.2% vs. 1.4%) before they reach the computationally expensive cross-modal transformer stage, since the AV-Deepfake1M++-calibrated thresholds are more conservative on InDeepFake's audio characteristics. This represents a favorable efficiency/accuracy trade-off rather than a free improvement: transfer reduces inference cost substantially while sacrificing under one point of AUC, which is a reasonable default when target-domain calibration data is unavailable, though native calibration remains preferable when such data can be obtained cheaply. On the fully silent AV-Deepfake1M++ subset, the gate correctly routes 100.0% of windows away from the synchronization pathway (AUC is undefined here since no synchronization score is computed), confirming that the routing mechanism degrades gracefully rather than producing spurious predictions on uninformative input.

G Statistical Grounding of the Synchronization Jitter Metric

We aggregate frame-level synchronization scores using S_{std}, motivated by the intuition that generative frame-blending introduces temporal instability in the synchronization signal. Here, we test this

intuition directly, ruling out two alternative explanations: that the instability is a boundary artifact of the segment stitching, and that S_{std} is a proxy for the number of edits present rather than for the quality of the underlying generative process.

Instability is intrinsic to generated content, not a stitching artifact. We partition manipulated windows into two groups: windows falling strictly within a manipulated segment, and windows adjacent to a manipulation boundary (where splicing artifacts, if present, would be expected to concentrate). We compare the local variance of the sync score, $\text{Var}_{t \in W}(s_t)$, between these groups. The manipulated-interior windows show higher local variance (mean STD $= 2.2296$) than boundary-adjacent windows (mean STD $= 1.8794$); a Mann-Whitney U test rejects the null hypothesis of equal distributions ($p < 0.001$). If temporal jitter were primarily a stitching artifact, we would expect the opposite pattern, with variance concentrated at segment boundaries rather than within them. This instead supports S_{std} as capturing a systemic failure of the generative model to hold its temporal latents stable throughout a manipulated segment — an intrinsic property of the forgery, not an easily-smoothed post-processing artifact.

Instability is independent of edit quantity. A separate concern is that S_{std} might simply increase with the number of manipulation segments in a video, rather than reflecting anything about generative quality per se. We compute the partial Spearman correlation between S_{std} and the number of active manipulated windows, controlling for total video duration. The resulting correlation is negligible ($r = 0.0010$, $p = 0.93$), indicating that S_{std} tracks the physical quality of the generative manifold rather than merely counting how many edits are present.

Relation to the embedding-space overlap in Appendix D. The partial overlap between real and fake samples visible in the t-SNE projection of Figure 5 (Appendix D) is consistent with, rather than in tension with, the analysis above: raw per-window sync scores alone provide an incomplete separation between real and fake content, which is precisely why aggregate temporal variance, rather than a raw score threshold, is necessary to disambiguate difficult samples. The local-variance and partial-correlation results in this section explain *why* that overlap exists (real and fake windows can have similar instantaneous alignment scores) and *why* S_{std} resolves it (the two classes differ systematically in how that alignment fluctuates over time, not in its instantaneous value).

H Computational Efficiency and Deployment Latency

To evaluate real-world deployment viability, we benchmark end-to-end wall-clock latency on a single NVIDIA H200 NVL GPU. Since feature extraction dominates the computational pipeline for multimodal deepfake detection, we report branch-specific overheads separately for full transparency, alongside two representative self-supervised baselines evaluated under identical conditions.

Table 8: End-to-end wall-clock latency for 5-second and 15-second input clips. "True Latency" includes feature extraction and classification; the primary pathway's breakdown (extraction + classification) is given in parentheses.

Architecture	True Latency (5s)	True Latency (15s)
Proposed AV Sync (Primary)	345.6 ms (345.0 ms extr. + 0.6 ms class.)	1,036.7 ms
Baseline AVH-Align	345.2 ms	1,035.2 ms
Baseline SpeechForensics	355.7 ms	1,067.0 ms
Proposed Worst-Case Fallback	618.4 ms (reuses audio extraction)	1,855.2 ms

During standard operation, the primary AV Sync pathway is highly efficient, processing a 5-second video in 345.6 ms; the classification step itself requires a negligible 0.6 ms, with the remainder spent on feature extraction. This operates well within real-time constraints and maintains latency parity with AVH-Align, a continuous-alignment baseline with no routing overhead, while offering superior discriminative performance.

We additionally analyze the worst-case computational scenario: a total synchronization failure requiring the system to route through the independent JEPA fallback. Because the architecture is

designed to reuse the audio features already extracted during the initial synchronization attempt, the fallback incurs only the residual cost of processing the visual features for the full-face stream. Consequently, the maximum latency remains bounded at 618.4 ms for a 5-second clip and under 2 seconds (1,855.2 ms) for a 15-second clip, confirming that the conditional-routing design provides robustness to missing or adversarial modalities without sacrificing real-time deployment speed.

I Ablation: The Mechanics of Aggregation and Context Length

Table 9: Ablation on temporal context length τ and test-time aggregation strategies. The standard deviation metric consistently dominates.

τ	Aggregation	AUC ↑	AP ↑
8	Mean	0.7379	0.8611
	5th Percentile	0.8141	0.9058
	Std	**0.8525**	**0.9434**
15	Mean	0.7424	0.8648
	5th Percentile	0.8099	0.9049
	Std	0.8463	0.9403
24	Mean	0.7409	0.8660
	5th Percentile	0.7990	0.9019
	Std	0.8239	0.9314

Table 9 consolidates the ablations for both aggregation strategy and temporal context length (τ). Mean pooling consistently collapses performance, confirming that localized partial deepfakes are washed out by global averaging. While low-percentile strategies are effective at isolating single catastrophic frame failures, the standard deviation metric dominates across all window sizes. Shorter temporal windows ($\tau = 8$) outperform broader contexts ($\tau = 15, 24$). Deepfake synchronization artifacts operate strictly at the millisecond boundary. Expanding τ incorporates excessive surrounding visual context, which dilutes the tightly coupled contrastive signal. This empirical degradation confirms that $\tau = 8$ provides the optimal receptive field for deepfake detection.