Do English Audio-Visual Deepfake Detectors Generalize to Indic Speech? Introducing SAMANVAYA.

Anonymous EMNLP submission

Abstract

Audio-Visual (AV) deepfake detection in Indic speech remains largely unaddressed. Existing detection systems are confined to high-resource English-centric data, and the artifacts they learn often fail to generalize under multilingual Indic distribution shifts, underscoring the challenges posed by Indic phonetic diversity and prosodic variability. We formalize this as the task of *Audio-Visual Indic Speech Deepfake Detection (AVISDD)* and introduce **SAMANVAYA**, a multimodal detection framework grounded in articulatory-acoustic temporal coupling–an intrinsic property of real speech production. Instantiating the *State Space Co-Evolution* mechanism, it learns a language and generator-stable signal: articulatory-acoustic coordination through explicitly modeling the temporal dynamics of lip-articulation and speech acoustics, and scores their cross-modal co-evolution to expose synthesis-induced desynchronization and dynamical inconsistencies. We benchmark existing English-trained detectors under (i) cross-lingual zero-shot transfer on the Indic multilingual InDeepFake dataset, (ii) In-distribution on the InDeepFake dataset, and (iii) transfer to the Hindi HAV-DF dataset. We observe severe performance collapse, evidencing distribution-local learning. **SAMANVAYA** achieves 0.99 AUC / 0.98 macro-F1 in-domain on InDeepFake and 0.92 AUC / 0.92 macro-F1 in zero-shot cross-dataset evaluation on HAV-DF, setting a new state-of-the-art for *AVISDD*.

1 Introduction and Background

Audio-Visual Deepfakes, synthetic videos in which a person's speech, appearance, or both are manipulated, are becoming a serious problem in recent times. This manipulation can contribute to impersonation, fraud, political disinformation, identity-based harm, and the creation of non-consensual explicit content. India has more than a billion people who speak over 120 languages and 19,500 dialects. India has one of the largest information consumption and hosts one of the world's most linguistically diverse and politically active media ecosystems, where a single manipulated video clip can be weaponized for misinformation, shaping public opinion, and influencing political narratives and electoral discourse(Hancock and Bailenson, 2021). Despite these vulnerabilities, a dedicated audio-visual deepfake detection system for Indic speech remains largely unaddressed.

AV speech follows a tightly coupled temporal relationship of acoustic signals and articulatory movements. In natural speech, lip movements and acoustic output are not independent; rather, there exists a structured synchrony between the lip movements and speech in which articulatory motion and acoustic speech evolve in a coordinated manner. This articulatory-acoustic temporal coupling is an intrinsic property of biological speech production and remains consistent across languages (Kadam et al., 2021). In contrast, AV speech deepfake generation pipelines often perform audio and visual synthesis in separate components, such as a text-to-speech system and a lip-sync model, and their post-fusion results do not replicate this coordination. This discrepancy leads to subtle misalignments in articulatory-acoustic temporal dynamics. Detecting such inconsistencies demands frameworks grounded in a signal that is stable across languages and generators, such as articulatory-acoustic temporal coupling, precisely because it is rooted in biology rather than in distributional artifacts.

Early approaches framed detection in videos purely as a vision problem, targeting visual manipulations such as DeepFaceLab (Perov et al., 2020) and Faceswap (Korshunova et al., 2017), (Rossler et al., 2019), or learning semantic face representations (Fang et al., 2025). With speech deepfakes emerging as an independent threat (Huang et al., 2025), they entered the audio-visual content space, fusing fake audio with real video, demanding frameworks such as SyncNet (Chung and Zisserman, 2016) and

LipForensics (Haliassos et al., 2021), which learn lip-audio synchrony and natural mouth motion representations. Further, modern deepfake pipelines are increasingly integrating synthetic speech and manipulated video by combining a TTS system with lip-sync models such as Wav2LIP (Prajwal et al., 2020), which demand joint audio-visual approaches (Shi et al., 2022), (Oorloff et al., 2024). (Datta et al., 2025) modeled phoneme-viseme associations through cross-attention-based pairwise matching between phonetic and visual speech representations on two English datasets. Considering the limitations of applying Western-centric approaches to Indic speech(Kumar et al., 2026), prior work has addressed Indic NLP resources (Kakwani et al., 2020; Khanuja et al., 2021; Ramesh et al., 2022) and multilingual synthetic speech (Sharma et al., 2025; Narayan et al., 2023), with limited effort towards AV speech deepfake detection (Kaur et al., 2024; Das et al., 2025). None of which addresses the architectural question of what detection signal generalizes across Indic languages and generators. On the architectural side, state space models (SSM) such as Mamba (Gu and Dao, 2023) have demonstrated their capacity to capture long-range temporal modeling in video understanding across various applications (Li et al., 2024; Mi et al., 2025), with the Mamba offering linear-time sequence modeling with a continuously evolving hidden state, properties well-suited to modeling the continuous coupled dynamics of articulatory-acoustic temporal coupling(Ruan et al., 2025).

Despite these advances, existing detectors share a fundamental limitation: they focus on distribution-local signals rather than properties that are stable across linguistic and generator distribution shifts. As (Smeu et al., 2025) states, video deepfake detectors take shortcuts, such as learning dataset-specific signals. The limitations are bipolar; first, they learn language-specific phoneme-viseme associations from English training data, i.e., phonological properties tied to the training language. Indic languages exhibit complex phonological properties absent in English (Girish et al., 2026). And these properties systematically disrupt the phoneme-viseme associations on which existing detectors rely(Agarwal et al., 2020). Secondly, and most importantly, even detectors trained on Indic data learn generator-specific artifacts, including visual inconsistencies, acoustic anomalies, and compression artifacts(Le-Phan et al., 2026). As a result, the failure is not merely a language-mismatch problem; models that perform well in in-distribution but collapse in cross-dataset settings, regardless of the language they were trained on, because their learned signal is distribution-local. No prior work has addressed this dual failure of cross-lingual and cross-dataset, nor has it provided a unified account of its cause beyond generic domain shift. The root of the dual failure is that existing approaches rely on distribution-specific artifacts, which capture features valid only within the training distribution. Due to the diversity of languages in the Indian context, what is needed is a stable detection signal that generalizes across languages and datasets. Articulatory-acoustic temporal coupling is precisely such a signal grounded in the dynamics of human biological speech and violated by all synthesis pipelines, regardless of the generator or language, since no pipeline that generates audio and visual streams independently can replicate the continuous, coupled dynamics of biological speech production. No prior work has designed an AV deepfake detection architecture grounded in this signal, nor has it deployed state-space models to capture cross-modal temporal co-evolution dynamics, the natural architectural choice for modeling the continuous, long-range, coupled evolution of articulatory and acoustic streams.

To address this, we introduce a multimodal framework **SAMANVAYA**, which means coordination and harmony in the Indic-Sanskrit language. As the name suggests, our framework measures whether audio and visual signals are in temporal coordination, as in real speech, or disrupted, as in deepfakes. Rather than learning distribution-local signals via attention, it measures whether the audio and visual hidden-state dynamics co-evolve consistently over time. To do so, we introduce a *State Space Co-Evolution* mechanism in **SAMANVAYA**, which processes audio and video through two parallel Mamba-based SSM streams, where each modality's temporal evolution is conditioned on the other via cross-modal gating. Also, to establish that the failure of existing detectors is structural rather than incidental, we conduct both cross-lingual zero-shot evaluation on InDeepFake and cross-dataset evaluation on HAV-DF. To the best of our knowledge, **SAMANVAYA** is the first audio-visual deepfake detection architecture explicitly designed for Indic speech, grounds detection in articulatory-acoustic temporal coupling as a distribution-stable signal, rather than in distribution-local frame-level corre-

lations.

Accordingly, our contributions are threefold:

- We introduce **SAMANVAYA**, the first dedicated audio-visual deepfake detection architecture for Indic speech that models temporal synchrony, built on a *Selective State Space Co-Evolution* mechanism, a cross-modally gated dual Mamba architecture that measures articulatory-acoustic co-evolution consistency as a detection signal.
- We conduct systematic cross-lingual (English $\rightarrow$ Indic) and cross-dataset (InDeepFake $\rightarrow$ HAV-DF) transfer evaluations of existing detectors, and provide evidence that their failure is attributed to the learning of distribution-local signals.
- We formally define the task *Audio-Visual Indic Speech Deepfake Detection (AVISDD)* and demonstrate via extensive experimentation that **SAMANVAYA** establishes state-of-the-art results on *AVISDD* across both in-distribution and cross-dataset settings.

We will release the code and the trained model after the double-blind review process.

2 Modeling

2.1 Feature Extraction

We select mHuBERT-147 (Boito et al., 2024) and VideoMAE (Tong et al., 2022) to obtain acoustic and articulatory motion representations from raw audio and the lip region of the video, respectively. mHuBERT-147 is a multilingual self-supervised speech model pre-trained on speech data of 147 languages. We chose this model for its multilingual capabilities, and its pretraining includes languages from the Indic families, enabling it to produce representations sensitive to Indic languages rather than monolingual English encoders, which might produce acoustic representations anchored to English phonological structure, undermining cross-lingual robustness. We deliberately chose VideoMAE over frame-level visual encoders as we intend to capture lip motion rather than instantaneous lip positions. VideoMAE is pretrained on a spatiotemporal objective, producing representations sensitive to motion patterns across frames, making it an appropriate visual encoder for our model, as its detection mechanism is grounded in articulatory dynamics rather than static appearance. Lip-region crops are extracted at 30 fps using RetinaFace (Deng et al., 2020), with bounding boxes centered on mouth corners and scaled by $1.5\times$ the inter-corner distance, and are temporally smoothed (Gaussian, $\sigma = 1.5$). All crops are resized to 224×224 and normalized with ImageNet statistics.

2.2 Proposed Framework: SAMANVAYA

We propose **SAMANVAYA**, a multimodal framework for *AVISDD* that models detection as a problem of temporal co-evolution between audio and visual streams. The overall pipeline is illustrated in Figure 1. Rather than learning distribution-local signals, it evaluates whether the internal dynamics of acoustic and articulatory representations evolve in a coordinated manner, arising from articulatory-acoustic temporal coupling, a fundamental property of speech production. Generative models that synthesize audio and video through separate systems often fail to preserve this co-evolution consistency and introduce subtle inconsistencies in the cross-modal temporal dynamics that **SAMANVAYA** is designed to detect. To operationalize this, **SAMANVAYA** introduces a *State Space Co-Evolution* mechanism that processes audio and video modalities in two parallel Mamba-based streams, where each modality's temporal evolution is co-modulated via cross-modal gating. The resulting representations are then used to make a decision through a synchrony-aware classification module.

Starting with the feature extraction as mentioned in 2.1, the extracted features of raw audio from mHuBERT-147: $\mathbf{X}^{(a)}_{\text{raw}} \in \mathbb{R}^{T_a \times 1024}$ where T_a denotes the number of mHuBERT-147 output frames, and the extracted features from the VideoMAE-Base of the lip crop sequence V, partitioned into non-overlapping clips of 16 consecutive frames is denoted as with the final clip padded by frame repetition if T is not divisible by 16, denoted as $\mathbf{X}^{(v)} = \text{Interp}\left(\{c^{(i)}\}_{i=1}^{\lceil T/16 \rceil}, T\right) \in \mathbb{R}^{T \times 768}$. Later the audio and video features are aligned by linearly interpolating the audio feature sequence from T_a frames to T frames then, $\mathbf{X}^{(a)} = \text{Interp}(\mathbf{X}^{(a)}_{\text{raw}}, T) \in \mathbb{R}^{T \times 1024}$. Both aligned sequences are projected to a shared space of dimension d_{model} using learned linear projections, followed by layer normalization: $\mathbf{H}^{(x)} = \text{LN}(\mathbf{X}^{(x)} W^{(x)})$, where $x \in a, v$ and $\mathbf{H}^{(x)} \in \mathbb{R}^{T \times d_{\text{model}}}$. We set $d_{\text{model}} = 512$ in all experiments. A single feature vector $h_t^{(m)}$ cannot capture articulatory dynamics, since phonological information depends on how lip movements evolve

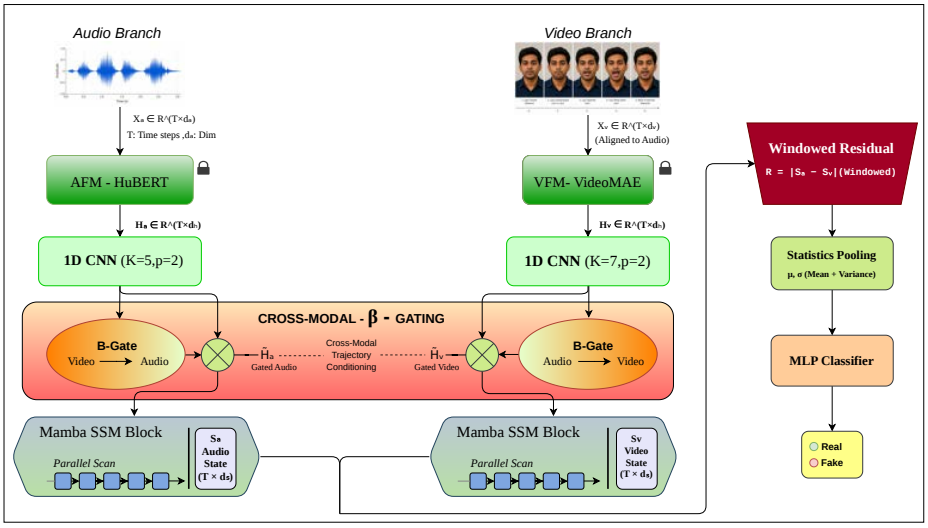


Figure 1: Overview of the proposed SAMANVAYA architecture.

over time rather than a static frame. Therefore, we apply depthwise 1D convolutions to capture local temporal motion before applying cross-modal gating: $\tilde{\mathbf{H}}^{(x)} = \mathrm{Conv1D}(\mathbf{H}^{(x)}, k_x),\ x \in \{a, v\}$, where $k_a = 5$ frames (≈167ms at 30fps) and $k_v = 7$ frames (≈233ms at 30fps), and $\tilde{\mathbf{H}}^{(a)}, \tilde{\mathbf{H}}^{(v)} \in \mathbb{R}^{T \times d_{\text{model}}}$. The asymmetric kernel sizes are chosen as acoustic dynamics evolve faster than the articulatory movements(Haliassos et al., 2021).

Cross-Modal B-Gate Co-Modulation. In the Mamba state-space framework, three time-varying parameter matrices govern the state dynamics at each step: A_t (state transition), B_t (input gate), and C_t (output projection). In standard unimodal Mamba, all parameters are computed from the same modality, but we modify only the input gate B_t to be conditioned on the other modality's trajectory, enabling cross-modal influence during state updates (Why the gate B_t? Refer Appendix C). For the audio stream, the input gate $B_t^{(a)} \in \mathbb{R}^{d_{\text{state}} \times d_{\text{model}}}$ is modulated by the video trajectory at time t, and vice-versa for the video stream:

$$\begin{aligned} B_t^{(a)} &= B^{(a)} \odot \sigma(W_{v \to a}\, \tilde{h}_t^{(v)}), \\ B_t^{(v)} &= B^{(v)} \odot \sigma(W_{a \to v}\, \tilde{h}_t^{(a)}) \end{aligned} \tag{1}$$

where $B^{(a)}, B^{(v)} \in \mathbb{R}^{d_{\text{state}} \times d_{\text{model}}}$ are base input gate matrices and $W_{v \to a}, W_{a \to v} \in \mathbb{R}^{d_{\text{state}} \times d_{\text{model}}}$ are cross-modal projections; $\sigma(\cdot)$ denotes the sigmoid and $\odot$ elementwise multiplication, with $\sigma(\cdot) \in (0, 1)$ scaling the gates without altering their structure. In real speech, both signals are generated from a single source, yielding consistent trajectories and stable gate modulations. In generated speech, audio and video signals are produced by separate systems with imperfect temporal coordination, resulting in inconsistent cross-modal trajectories that produce conflicting gate modulations and divergent hidden-state dynamics. Since $B_t^{(a)}$ and $B_t^{(v)}$ are functions exclusively of the pre-computed trajectory vectors $\tilde{\mathbf{H}}^{(a)}$ and $\tilde{\mathbf{H}}^{(v)}$, all gate matrices for the complete sequence $t = 1, \ldots, T$ are computed in a single parallel forward pass prior to scanning, preserving the full parallelizability of the associative scan and maintaining the linear time complexity of Mamba.

Parallel Selective State Space Scan. Having the gate matrices $\{B_t^{(a)}\}_{t=1}^{T}$ and $\{B_t^{(v)}\}_{t=1}^{T}$ precomputed, two independent Mamba SSM streams process the audio and video sequences in parallel. The hidden states evolve over time via a standardized discrete-time recurrence, obtained via zero-order hold discretization of the continuous-time state space model:

$$\begin{aligned} s_{t+1}^{(a)} &= \bar{A}^{(a)} s_t^{(a)} + B_t^{(a)} h_t^{(a)}, \\ s_{t+1}^{(v)} &= \bar{A}^{(v)} s_t^{(v)} + B_t^{(v)} h_t^{(v)} \end{aligned} \tag{2}$$

where $s_t^{(a)}, s_t^{(v)} \in \mathbb{R}^{d_{\text{state}}}$ are the hidden states at time t, $\bar{A}^{(a)}, \bar{A}^{(v)} \in \mathbb{R}^{d_{\text{state}} \times d_{\text{state}}}$ are learned diagonal state transition matrices after discretization, $h_t^{(a)}, h_t^{(v)} \in \mathbb{R}^{d_{\text{model}}}$ are the t-th rows of $\mathbf{H}^{(a)}$ and $\mathbf{H}^{(v)}$ respectively, and both streams are initialized from the zero state $s_1^{(a)} = s_1^{(v)} = \mathbf{0}$. The streams share parameters only through $W_{v \to a}$ and $W_{a \to v}$. The diagonal, time-invariant $\bar{A}$ ensures stable dynamics while preserving parallelism, as input-dependent transitions would introduce sequential dependencies. The recurrences are com-

puted using parallel associative scan, achieving $O(T)$ complexity versus $O(T^2)$ for cross-attention, producing two hidden state sequences:

$$\begin{aligned} \mathbf{S}^{(a)} &= \{s_t^{(a)}\}_{t=1}^{T} \in \mathbb{R}^{T \times d_{\text{state}}}, \\ \mathbf{S}^{(v)} &= \{s_t^{(v)}\}_{t=1}^{T} \in \mathbb{R}^{T \times d_{\text{state}}} \end{aligned} \tag{3}$$

Windowed Residual Computation. The hidden states $s_t^{(a)}$ and $s_t^{(v)}$ encode temporal dynamics of acoustic and articulatory streams. In generated speech, the imperfect post-hoc coordination of separate generative systems for audio and visual introduces inconsistencies that are seen as divergence between $s_t^{(a)}$ and $s_t^{(v)}$ in a shared comparison space. Let $\phi, \psi : \mathbb{R}^{d_{\text{state}}} \to \mathbb{R}^{d_{\text{proj}}}$ be learned projections to a shared space ($d_{\text{proj}} = 256$), followed by L2 normalization to map representations onto the unit hypersphere $\mathcal{S}^{d_{\text{proj}}-1}$.

$$\hat{s}_t^{(a)} = \frac{\phi(s_t^{(a)})}{\|\phi(s_t^{(a)})\|_2 + \epsilon}, \; \hat{s}_t^{(v)} = \frac{\psi(s_t^{(v)})}{\|\psi(s_t^{(v)})\|_2 + \epsilon} \tag{4}$$

where $\epsilon = 10^{-8}$ ensures numerical stability and $\hat{s}_t^{(a)}, \hat{s}t^{(v)} \in \mathcal{S}^{d\text{proj}-1}$. L2 normalization prevents collapse of ϕ and ψ, and bounds distances in $[0, 4]$, making them consistent with cosine similarity. Instead of forcing exact frame-by-frame alignment, we compare each video state with nearby audio state with a small temporal window and take the closest match. In natural speech, lip movements often occur slightly before the corresponding audio (around 50–120 ms)(Schwartz and Savariaux, 2014), so a soft-min is used to approximate the minimum distance while keeping the formulation differentiable.

$$r_t = -\tau \log \sum_{k=-w}^{w} \exp\left(-\frac{\|\hat{s}_{t+k}^{(a)} - \hat{s}_t^{(v)}\|_2^2}{\tau} \right) \tag{5}$$

where $\tau > 0$ is a temperature parameter and boundary indices are clamped to $[1, T]$. As $\tau \to 0$ converges to a hard minimum, and $\tau \to \infty$ to a constant independent of the distances. We set $w = 4$ frames, covering the 50–120 ms coarticulation range. The temperature τ is annealed from 0.1 to 0.01 using a cosine schedule. Small r_t indicates consistent co-evolution, while large r_t reflects audio-visual mismatch.

Aggregation and Classification. Then mean, and variance pooling is applied to the residual sequence $\mathbf{r} = r_{t\,t=1}^{T}$, and a fixed-dimensional vector is obtained. The mean captures the synchrony quality across the full video, and the variance captures the temporal consistency of synchrony. The deepfakes may exhibit localized co-evolution spikes, producing elevated variance even when the mean is moderate. The feature vector is passed through a two-layer fully connected network with 256 units, GELU activation, and dropout ($p = 0.3$), followed by a single output unit with sigmoid activation to obtain class probabilities, and optimized using binary cross-entropy loss over the training set.

3 Experiments and Results

3.1 Datasets

InDeepFake(Das et al., 2025): It is the first comprehensive multilingual and multimodal dataset for Indian deepfakes, consisting of over 4600 high-quality videos, including baseline footage and synthetic deepfakes. The dataset strictly limits its manipulations to faces and speech. It covers seven distinct Indic languages and three major audio-visual manipulations: i) Face-Swapping and Re-enactment (Visual) via FaceSwap, DeepFaceLab, FSGAN (Face Swapping GAN), ii) Lip-Syncing (Audio-Visual Alignment) via Wav2Lip, and iii) Voice Cloning / Text-to-Speech (Audio) via SV2TTS, E2TTS, and F5TTS models. To rigorously evaluate the proposed architecture, we establish InDeepFake as our primary experimental dataset.

Hav-DF(Kaur et al., 2024): It is the first large-scale multimodal Hindi-language deepfake dataset, consisting of 508 videos (200 authentic, 308 manipulated), where manipulations were supported by face-swapping, lip-syncing, and voice cloning via models such as ReTalking, RVC, and FSGAN. It serves as a rigorous, Hindi-specific out-of-distribution (OOD) benchmark for evaluating cross-generator and cross-dataset generalization capabilities in Indic speech. We use this dataset to test the transferability of InDeepFake-trained detectors on the unseen Indic-Hindi language.

FakeAVCeleb(Khalid et al., 2021): It is one of the largest multimodal English deepfake datasets, comprising over 20000 videos curated from the VoxCeleb2 YouTube corpus, and manipulations include Real/Fake Audio and Video combinations. The dataset is restricted entirely to English speech, serving as the standard high-resource western-centric multimodal benchmark.

Model	InDeepFake		InDeepFake → HAV-DF		Δ (Drop)	
	AUC	M-F1	AUC	M-F1	AUC	M-F1
AVFF (Oorloff et al., 2024)	0.94	0.91	0.88	0.47	-0.06	-0.44
ConLLM (Kashyap et al., 2026)	0.95	0.94	0.83	0.52	-0.12	-0.42
PIA (Cross-Attention) (Datta et al., 2025)	0.91	0.90	0.80	0.44	-0.11	-0.46
UniForensics (Fang et al., 2025)	0.87	0.84	0.55	0.61	-0.32	-0.23
LipForensics (Haliassos et al., 2021)	0.86	0.83	0.59	0.55	-0.27	-0.28
SyncNet (Chung and Zisserman, 2016)	0.82	0.76	0.61	0.50	-0.21	-0.26
AV-HuBERT (Shi et al., 2022)	0.92	0.89	0.83	0.55	-0.09	-0.34
Samanvaya (Proposed)	**0.99**	**0.98**	**0.92**	**0.92**	**-0.07**	**-0.06**

Table 1: In-distribution and cross-dataset performance Smaller drop (Δ) indicates better generalization.

3.2 Training Details

For cross-lingual zero-shot evaluation of existing English detectors, models were trained on the FakeAVCeleb dataset and tested on the IndeepFake corpus. For in-distribution evaluation of **SAMANVAYA**, we used a $70/10/20$ stratified split of IndeepFake for training, validation, and testing, averaged over 5 runs with different random seeds. To test cross-dataset generalization on unseen Indic speech, the models trained on IndeepFake are evaluated in zero-shot mode on HAV-DF. In ablation studies (removing the CNN block, cross-modal gating, SSM block and, residual module), we followed the same $70/10/20$ split and evaluation protocol. We report AUC and macro-F1 as evaluation metrics. We follow a zero-intervention policy for all baselines and their respective modalities. Please refer to Appendix B and F for detailed information on baselines and training and architecture hyperparameters.

3.3 Results and Discussion

Our experimentation spans across four settings: i) zero-shot evaluation of existing English-trained detectors on multilingual Indic speech, ii) in-distribution evaluation on InDeepFake, iii) a cross-dataset evaluation to measure generalization across Indic datasets, and iv) ablation studies to analyze the contribution of key components.

Model	AUC	macro-F1
AVFF	0.90	0.45
ConLLM	0.91	0.51
PIA (Cross-Attention)	0.86	0.48
UniForensics	0.53	0.48
LipForensics	0.57	0.53
AV-HuBERT	0.90	0.52
SyncNet	0.62	0.55

Table 2: Zero-shot evaluation of English-trained detectors on multilingual Indic speech (InDeepFake).

Variant	AUC	macro-F1
Full Model	**0.99**	**0.98**
w/o Cross-modal Gating	0.94	0.91
w/o SSM (Mamba)	0.93	0.91
w/o CNN (Trajectory)	0.96	0.93
w/o Residual Module	0.91	0.90

Table 3: Ablation study showing the contribution of each component of Samanvaya.

Table 2 presents the zero-shot performance of seven English-trained detectors when evaluated on the InDeepFake dataset. Looking at the AUC scores, the numbers are reasonably impressive ($0.62 - -0.91$), indicating that models hold the ability to detect the real and fake samples correctly. But the macro-F1 scores collapse substantially ($0.45 - -0.55$), indicating that the models fail to make calibrated binary decisions on the new distribution. Inspection of confusion matrices (Appendix E) shows a degenerate prediction behavior: models trained on English data default to classifying nearly every Indic sample as fake, resulting in a near-zero true negative rate. This pattern indicates a threshold miscalibration under domain-shift (Ovadia et al., 2019). This failure is not an indication to poor model design or its inability to learn the signal. As shown in Table 1, models such as AVFF (AUC 0.90), ConLLM (AUC 0.91), and AV-HuBERT (AUC 0.90) demonstrate strong performance in in-distribution settings on Indic data. The failure is in generalization, and this can be attributed to the learning of dataset-specific artifacts, such as visual artifacts, acoustic anomalies, and generator-induced patterns from the English data. Consequently, the distribution of such artifacts shifts, as in our experiments from English to Indic, the models no longer efficiently separate the real and fake classes and collapse to a majority prediction, treating anything not from the learned

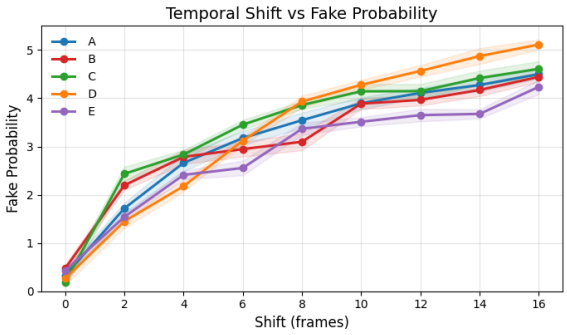
artifacts as a fake sample. Visual-oriented models further degrade. Lip-Forensics and Uniforensics achieve AUCs of $0.53 - 0.57$ and macro-F1 of $0.48 - 0.53$, indicating their inability to either rank or classify well on Indic data. Even SyncNet, which is more sensitive to temporal-based relationships achieves near-random performance in zero-shot cross-lingual transfer. This uniformity of failure of such varying paradigms in zero-shot cross-lingual transfer to Indic speech is the central finding of this experiment. Models that rely on distribution-local artifacts cannot generalize to multilingual Indic speech. This motivates us to develop an Indic-specific framework designed to handle the multilingual and phonologically diverse properties of Indic languages.

To assess the performance of existing detectors and our proposed framework in Indic multilingual languages, we perform an in-distribution evaluation on the InDeepFake dataset, and Table 1 reports the performance of all models. **SAMANVAYA** achieves the highest performance across all models, yet the nearest competitors are ConLLM and AVFF, which achieve 0.95 and 0.94 AUC, and 0.94 and 0.91 macro-F1, respectively. Compared to the cross-attention baseline, the 8-point gap directly assesses the effectiveness of *Selective State Space Co-evolution* in our framework. We observed an AUC score of 0.99 for our framework and scrutinized it for potential overfitting to distribution-specific artifacts; the subsequent cross-dataset evaluation on the HAV-DF supports generalization rather than dataset memorization.

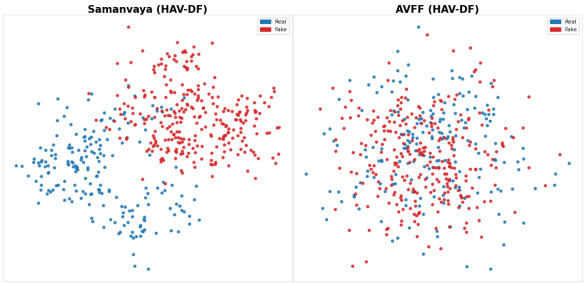
To further assess how these multilingual Indic-trained models on InDeepFake perform on an unseen Indic dataset, we conduct a cross-dataset zero-shot evaluation on the HAV-DF, an Indic Hindi-language AV speech deepfake dataset. Table 1 reports the corresponding evaluation. **SAMANVAYA** achieves 0.92 AUC and 0.92 macro-F1, degrading by only 7 AUC points and 6 macro-F1 points from its Indeepfake In-distribution performance. Every other model degrades substantially. The degradation pattern among the baselines reproduces and extends the threshold miscalibration failure observed in Table 2. AVFF, which achieves 0.91 macro-F1 on InDeepFake, drops to 0.47 on HAV-DF, a loss of 0.44 points. ConLLM drops from 0.94 to 0.52, a loss of 0.42 points. The cross-attention baseline drops from 0.90 to 0.44, a loss of 0.46 points. AV-HuBERT drops from 0.89 to 0.55, a loss of 0.34 points. The visual-oriented models Lipforensics and Uniforensics follow the same degradation pattern as we discussed in Table 2. Even a synchrony-sensitive model, SyncNet, was unable to transfer across Indic speech data.

Connecting failure patterns across evaluation settings reveals a consistent, structured generalization deficit. Under-zero-shot transfer from English to multilingual Indic speech data: all models collapsed, exhibiting model-specific behaviors. The pronounced divergence (Δ) between AUC and macro-F1 scores is indicative of systematic threshold miscalibration and a fundamental inability to transfer across linguistic boundaries. This motivates Indic-specific modeling. However, training on Indic data alone does not resolve this issue. The same baselines, trained and evaluated in-distribution on the multilingual Indic dataset InDeepFake, achieve strong performance; however, under zero-shot cross-dataset evaluation on HAV-DF, the same failure pattern resurfaces. The dual failure confirms that models learn distribution-local artifacts rather than stable generator-agnostic representations of authentic speech. This further strengthened the generalization failure, and it is not reducible to language mismatch. It cannot be resolved solely by training on Indic data. *Training on Indic speech is necessary but insufficient;* robust detection requires a signal grounded in properties that are invariant across both languages and generation pipelines. Our proposed framework **SAMANVAYA** addresses this directly, achieving state-of-the-art performance in In-distribution settings and, critically, under zero-shot cross-dataset transfer on HAV-DF, establishing it as the most generalizable framework for *AVISDD*.

Why SAMANVAYA generalizes? The generalization advantage of **SAMANVAYA** is a direct consequence of what the model is trained to measure. Natural speech is a coupled process of articulatory-acoustic dynamics, and their temporal dynamics are inherently coordinated. Synthetic speech breaks this coupling because it originates from separate sources and is approximated as coordinated at best. It's *Selective State Space Co-Evolution* mechanism is built around this distinction. The cross-modal B-gate modulation modulates each stream's memory update on the current dynamics of the other; as a result, the hidden states of audio-visual streams co-evolve in mutual influence rather than in isolation. It's like the Video



(a) Temporal shift sensitivity analysis on HAV-DF.



(b) t-SNE visualization of learned representations on HAV-DF.

stream tells the Audio stream: "Hey, I'm seeing a 'P' sound being formed by the lips, you should probably listen for a 'P' sound now." The Mamba-based parallel associative scan propagates this mutual influence across the full extent of the sequence, capturing long-range dependencies that frame-level matching cannot. The windowed residual then measures whether the two streams' internal representations are in close alignment (i.e., genuine speech co-production) or divergent (i.e., synthetic speech production). It remains tied to whether both streams are produced by a single source and co-evolve; this property holds across all datasets and languages. That is why the threshold **SAMANVAYA** learns on InDeepFake remains valid on HAV-DF, while every other evaluated model's threshold collapses.

The Figure 2a offers direct evidence of what **SAMANVAYA** is sensitive to. 5 Real samples were taken from HAV-DF as is, the video frames remain untouched, and audio frames are shifted against it in the range of 0 to 16. As the temporal offset grows, the model's probability of fakeness rises sharply. No label was changed, no synthesis was involved, the only change was coordination. The model flags desynchronized real speech as manipulated, and this is not a failure, it confirms that the detection signal is grounded in cross-modal temporal coherence rather than in any acoustic or visual artifact. And Figure 2b extends this picture, showing the t-SNE plot for fused representations of audio and video sequences, from our model and AVFF (the nearest baseline in-distribution) under cross-dataset performance on HAV-DF. **SAMANVAYA**'s representations form well-separated clusters, whereas those of AVFF collapse into an entangled distribution. Furthermore, the Ablation Study 3 strengthens this claim by showcasing the contribution of each component. Please refer the Appendix D, and Table 3.

4 Conclusion

In this work, we initiated a focused study on Audio-Visual Deepfake Detection on multilingual Indic speech. Through cross-lingual zero-shot evaluation on the InDeepFake dataset, we observe that existing models collapse, demonstrating their inability to generalize to Indic speech. Supporting this, we introduced **SAMANVAYA**, a co-evolution-driven framework that models articulatory-acoustic temporal coordination via cross-modal selective state-space dynamics. Assessing its performance, we conducted an in-distribution evaluation on the InDeepFake dataset, followed by a cross-dataset zero-shot evaluation on the Hindi HAV-DF dataset. Existing assessments assessed good results in in-distribution settings but collapsed further under the cross-dataset evaluation. The dual failure validated the generalization weaknesses of English-trained models under language and distribution shifts. The Δ (Drop) in AUC and macro-F1 in zero-shot cross-lingual, the Δ (Drop) in in- vs out-of-distribution on HAV-DF, and the Δ (Drop) in AUC and macro-F1 on HAV-DF transfer evidence that models learn distribution-local artifacts rather than language or generator-stable signals. Beyond the immediate contribution, these findings point towards a broader principle: signals grounded in the intrinsic properties of real speech generalize, whereas those grounded in observable properties of fake content do not. Owing to this, **SAMANVAYA**'s performance in in-distribution on InDeepFake (0.99 AUC / 0.98 macro-F1), and in zero-shot cross-dataset evaluation on HAV-DF (0.92 AUC / 0.92 macro-F1), establishes a new state-of-the-art for Audio-Visual Indic Speech Deepfake Detection *AVISDD*.

5 Limitations.

Here we discuss several limitations that need acknowledgment. At first, InDeepFake provides no language-specific labels, which prevents per-language performance analysis, even though the dataset spans multiple Indic languages, which lack an extensive evaluation to characterize per-language behavior. Second, cross-dataset generalization is validated on a single Hindi target dataset, and an extension to other Indic languages and generation methods remain open. Third, both encoders are frozen, and their representations are used for further modeling, though full finetuning can be explored given the increased model size and inference time. Finally, this work remains focused on Indic languages, examining only their phonological diversity across languages and dialects, and lacks a transferability evaluation to other phonologically distinct languages, such as Southeast Asian languages.

6 Ethical Considerations

While audio-visual deepfake detection systems can help mitigate the risks of misinformation, impersonation, and non-consensual synthetic media, they also carry the potential for false positives. Our work is intended solely for research purposes and as an effort to improve the detection of deepfakes in multilingual Indic speech. We emphasize that such system should be deployed under human supervision and not be considered a fully reliable automated system for authenticating media. We adhere to the original dataset and model licenses documented in their respective works and provide appropriate citations to all datasets, pretrained models, and baseline frameworks used in this study.

7 AI Assistance Disclosure

AI-based writing assistants were used for language refinement, editing, and polishing. All scientific claims, experimental design, implementation, and final content decisions were made and verified by the authors.

References

Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. 2020. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mhubert-147: A compact multilingual hubert model. *arXiv preprint arXiv:2406.06371*.

Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer.

Arnab Kumar Das, Aritra Bose, Priya Manohar, Anurag Dutta, Ruchira Naskar, and Rajat Subhra Chakraborty. 2025. Indeepfake: A novel multimodal multilingual indian deepfake video dataset. *Pattern Recognition Letters*.

Soumyya Kanti Datta, Tanvi Ranga, Chengzhe Sun, and Siwei Lyu. 2025. Pia: Deepfake detection using phoneme-temporal and identity-dynamic analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1596–1606.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212.

Ziyuan Fang, Hanqing Zhao, Tianyi Wei, Wenbo Zhou, Ming Wan, Zhanyi Wang, Weiming Zhang, and Nenghai Yu. 2025. Uniforensics: Face forgery detection via general facial representation. *IEEE Transactions on Dependable and Secure Computing*.

Girish, Mohd Mujtaba Akhtar, Orchid Chetia Phukan, and Arun Balaji Buduru. 2026. Indic-codecfake meets satyam: Towards detecting neural audio codec synthesized speech deepfakes in indic languages.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. 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*, pages 5039–5049.

Jeffrey T. Hancock and Jeremy N. Bailenson. 2021. The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3):149–152.

Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. 2025. Speechfake: A large-scale multilingual speech deepfake dataset incorporating cutting-edge generation methods. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9985–9998.

Anup Kadam, Sagar Rane, Arpit Mishra, Shailesh Sahu, Shubham Singh, and Shivam Pathak. 2021. A survey of audio synthesis and lip-syncing for synthetic video generation. *EAI Endorsed Transactions on Creative Technologies*, 8(28):169187.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Gautam Siddharth Kashyap, Harsh Joshi, Niharika Jain, Ebad Shabbir, Jiechao Gao, Nipun Joshi, and Usman Naseem. 2026. Revealing the truth with conllm for detecting multi-modal deepfakes. *arXiv preprint arXiv:2601.17530*.

Sukhandeep Kaur, Mubashir Buhari, Naman Khandelwal, Priyansh Tyagi, and Kiran Sharma. 2024. Hindi audio-video-deepfake (hav-df): A hindi language-based audio-video deepfake dataset. *arXiv preprint arXiv:2411.15457*.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. 2021. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2017. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685.

Raghvendra Kumar, Devankar Raj, and Sriparna Saha. 2026. Bhashasutra: A task-centric unified survey of indian nlp datasets, corpora, and resources. *arXiv preprint arXiv:2604.18423*.

Minh-Khoa Le-Phan, Minh-Hoang Le, Trong-Le Do, and Minh-Triet Tran. 2026. Robust deepfake detection: Mitigating spatial attention drift via calibrated complementary ensembles. *arXiv preprint arXiv:2604.25889*.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pages 237–255. Springer.

Yachun Mi, Yu Li, Weicheng Meng, Chaofeng Chen, Chen Hui, and Shaohui Liu. 2025. Mvqa: Mamba with unified sampling for efficient video quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18498–18509.

Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. 2023. Df-platter: Multi-face heterogeneous deepfake dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748.

Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. 2024. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, et al. 2020. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. 2025. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications*.

jean-luc Schwartz and Christophe Savariaux. 2014. No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS computational biology*, 10:e1003743.

Divya V Sharma, Vijval Ekbote, and Anubha Gupta. 2025. Indicsynth: A large-scale multilingual synthetic speech dataset for low-resource indian languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22037–22060.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. 2025. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18815–18825.

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

A Appendix Description

In this appendix, we provide the following:

- B: Detailed information about the existing baseline models used in the experimentation.
- C: Here we discuss the reason for selecting the B gate rather than A and C to condition the audio and visual stream before the SSM processing.
- D: Here, we present the ablation study in detail, discussing the importance of each component in our architecture and its contribution to the desired overall performance.
- F: Here in this section, we report the training and architectural hyperparameters of the proposed architecture.
- E: Here, we report the confusion matrices of our proposed architecture with the other latter three baseline models in the In-distribution and cross-dataset settings.

B Baseline Description

- AVFF (Oorloff et al., 2024): It is a two-stage deepfake-detection framework that combines self-supervised representational learning along with contrastive feature fusion to learn rich cross-modal representations. We include it in our experiment as a representative audio-visual feature fusion baseline, given its strong performance on the FakeAVCeleb dataset.
- ConLLM (Kashyap et al., 2026): It is a hybrid audio-visual deepfake detection framework that combines PTM-based representation extraction, aligning embeddings using contrastive learning, and leveraging LLM-guided reasoning to capture semantic inconsistencies. We include it as a large-model-based multimodal reasoning baseline and its state-of-the-art performance on the FakeAVCeleb dataset.
- PIA (Cross-Attention) (Datta et al., 2025): PIA is an audio-visual deepfake-detection framework that combines phoneme representations, lip geometry features, and facial identity embeddings through cross-attention-based alignment. We use PIA as it is the most closely related temporal-correspondence baseline, and also to evaluate whether frame-level phoneme-viseme alignment generalizes to Indic audio-visual deepfake detection.
- LipForensics (Haliassos et al., 2021): It is a visual-speech deepfake detection framework that models temporal irregularities in natural mouth movements, which are observed in manipulated videos. We include it in our baseline as a lip-motion-based unimodal baseline to evaluate whether visual articulatory dynamics alone are sufficient for deepfake detection.
- UniForensics (Fang et al., 2025): It is a unimodal deepfake-detection framework for videos. It leverages a facial encoder to understand the facial features, such as identity, geometry, and expressions, and a spatio-temporal transformer to model relationships between frames. We include it as a strong visual artifact and facial dynamics baseline for comparison against audio-visual approaches.
- AV-HuBERT (Shi et al., 2022): It is a self-supervised audio-visual representation learning framework that jointly models visuals and speech. In our experiments, we leverage a downstream pipeline utilizing AV-HuBERT. Here, visuals and audio are passed through it to obtain the embeddings, which are then temporally pooled and passed to the downstream classification head for real/fake prediction. We include it as a strong, pretrained audio-visual representation-learning baseline.
- SyncNet (Chung and Zisserman, 2016): It is an audio-visual deepfake detection framework that detects inconsistencies via measuring temporal correspondence between lip movements

and speech signals through contrastive synchronization learning. In our implementation, the lip and audio frame sequences are passed through SyncNet to obtain embeddings, which are then fed to downstream classification for real/fake prediction. We include it as a synchrony-based baseline to evaluate the robustness of alignment-sensitive detection

C Why we chose to modulate the B gate instead of the A and C gates?

We modulate B_t exclusively and not A_t or C_t for the following reasons. The A_t matrix works for intrinsic state decay, that is, the rate at which each stream forgets its accumulated history (forget gate). Acoustic phonological state and the visual articulatory state operates at different natural timescales, so their decay timing is inconsistent with one another, and in order to modulate them, they must retain modality-specific forgetting dynamics; cross-modal modulation of A_t would conflate these independent temporal processes. The C_t matrix is a post-hoc output projection runs after the scan (output gate), not during it. Modulating C_t cross-modally is functionally equivalent to late fusion by another name, and does not reach the encoding stage, so it can't be claimed as a contribution to how representations are built. The B_t matrix, as the input gate, decides what information from the current timestep enters the hidden state. This makes it the perfect location for the cross-modal influence, not the A_t or C_t. Since $B_t^{(a)}$ and $B_t^{(v)}$ are functions exclusively of the pre-computed trajectory vectors $\tilde{\mathbf{H}}^{(a)}$ and $\tilde{\mathbf{H}}^{(v)}$, all gate matrices for the complete sequence $t = 1, \ldots, T$ are computed in a single parallel forward pass prior to scanning, preserving the full parallelizability of the associative scan and maintaining the linear time complexity of Mamba.

D Ablation Study

Table 3 analyzes the contribution of each component of **SAMANVAYA** for *AVISDD*. The full model achieves 0.99 AUC / 0.98 macro-F1, and every ablation leads to a clear degradation, demonstrating that **SAMANVAYA** benefits the best from the joint contribution of Cross-modal gating, Selective State-space model Mamba, CNN block, and the Residual Module. Removing the residual module produces the largest single-component drop, reducing performance to 0.91 AUC and 0.90 macro-F1. This module is the direct unit that measures cross-modal hidden-state consistency, and its removal means the model loses its primary detection signal entirely and falls back to what discriminative information the MLP classifier can extract from the incoming audio-visual stream from SSM (Mamba). Without it the model is learning a signal but not merely the temporal consistency we meant it to learn. Removing the cross-modal B-gate modulation and passing the representations from the CNN block to SSM Mamba drops the performance to AUC 0.93. Without it, both the Mamba-streams are processing the audio and visual streams without them being aware of each other during state evolution. So the hidden states that reach the residual module are not mutually aware of each other; therefore, computed without knowledge of the opposing modality's current dynamics, reducing the co-evolution signal to a post-hoc comparison of independently evolved representations rather than representations shaped by active mutual conditioning. The performance drop evidences the meaningful contribution of mutual conditioning; the residual module shows greater consistency when the hidden states are evolved under cross-modal influence. Removing the SSM Mamba and replacing it with the Transformer encoder further degrades the performance to 0.93 AUC and 0.91 macro-F1. Here the gated features are passed through a standard transformer encoder. The drop demonstrates that temporal modeling is necessary beyond cross-modal gating, and that the sequential hidden-state evolution of the associative scan captures temporal dependencies in the audio-visual dynamics that the transformer cannot. Removing the CNN module shows the smallest drop among all ablation variants at 3 AUC points and 5 macro-F1 points, which was the expected result. It provides motion-aware context for cross-modal gating, but the gating mechanism can be controlled by the Encoders' instantaneous representations. Taken together, each component shows its contribution to modeling temporal audio-visual dynamics and, ultimately, to capturing the generalized detection signal.

E Consfusion Matrices

Here, we visually demonstrate the confusion matrices for **SAMANVAYA** and the top three models in the in-distribution and cross-dataset settings, highlighting the comparison and evidence of dual failure. Figure 3 reports the confusion matrices of

SAMANVAYA, AVFF, ConLLM, and PIA under the In-distribution evaluation on InDeepFake and cross-dataset evaluation on the HAV-DF dataset. We chose only the above models to showcase the confusion matrices, as they deliver comparable performance to the proposed framework in In-distribution settings, yet collapse in zero-shot cross-dataset HAV-DF. The confusion matrices evidence the collapse and distribution-local learning of existing English-trained models and the generalization of our proposed framework.

F Training Details and Hyperparameters

Table 4 lists the full set of training and architectural **SAMANVAYA** hyperparameters used across all experiments. All experiments are performed on NVIDIA A100 SXM4 40GB workstation using PyTorch 2.3.1, CUDA 12.2, and Python 3.10.

Hyperparameter	**Value**
Audio Encoder	mHuBERT-147 (Frozen)
Video Encoder	VideoMAE-Base (Frozen)
Audio Feature Dim	$d_a = 1024$
Video Feature Dim	$d_v = 768$
Projected Feature Dim	$d_model = 512$
SSM State Dim	$d_state = 64$
Residual Projection Dim	$d_{\text{proj}} = 256$
Conv1D for Audio	$k_a = 5$
Conv1D for Video	$k_v = 7$
SSM Backbone	Mamba
SSM Layers	4
Residual Pooling	Mean + Variance
Optimizer	AdamW
AdamW Betas	$(0.9, 0.999)$
AdamW Epsilon	10^{-8}
Weight Decay	0.01
Learning Rate	1×10^{-4}
Gradient Clipping	1.0
Dropout	0.3
Batch Size	16
Epochs	50
Warmup Ratio	10%
Scheduler	Cosine Decay
Loss Function	Binary Cross-Entropy

Table 4: Training and architectural hyperparameters used for SAMANVAYA across all experiments.

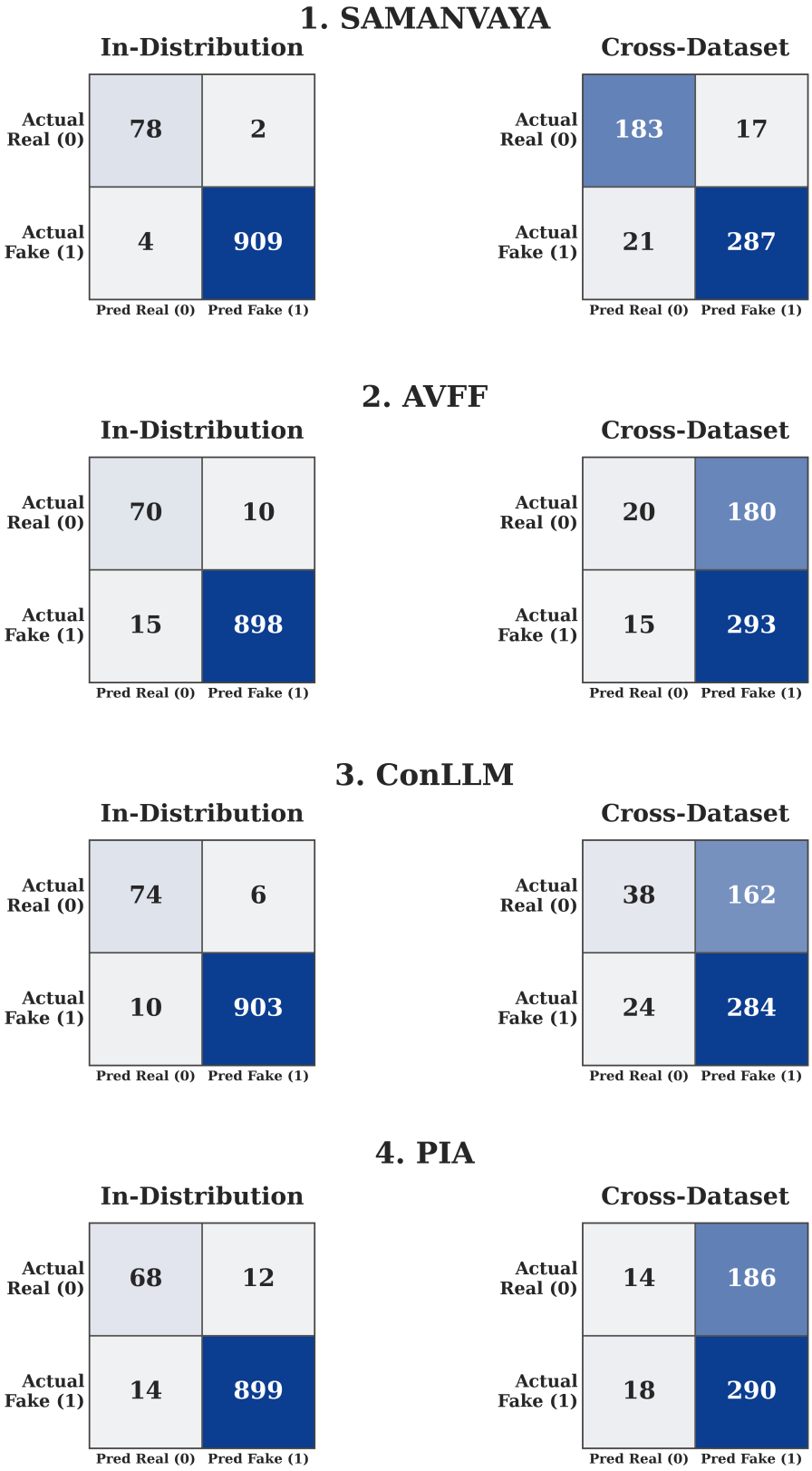


Figure 3: Comparing the confusion matrices of the top 3 baselines with our proposed framework.