SAMVAD-DF: UNMASKING PARTIAL AUDIO DEEPFAKES IN MULTISPEAKER CONVERSATIONAL SETTINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio deepfake detection is entering a new regime. Rather than synthesizing entire utterances, emerging attacks manipulate authentic recordings by seamlessly replacing a sentence or even a single word. While visual deepfake research has progressively expanded from isolated face swaps to complex multi-person scenarios, audio detection has largely remained focused on single-speaker English speech and binary utterance-level classification, with limited emphasis on localizing manipulated content. To address this gap, we introduce SAMVAD-DF, the first large-scale benchmark for partial audio deepfakes in multispeaker Hindi conversations. Built from 107 hours of in-the-wild debates and podcasts, the dataset captures substantial acoustic diversity, containing recordings with up to 76 speakers. SAMVAD-DFdefines three granularities of manipulation, ranging from speaker-level manipulations to highly challenging sub-second word-level replacements. Constructing this benchmark also reveals a fundamental limitation of conventional training. Existing pipelines assign file-level labels to every short audio window, causing clean regions far from the splice to be treated as manipulated during supervision. We instead introduce a validated, strictly bounded temporal annotation protocol that labels only the true manipulated span, enabling faithful localization learning. Experiments demonstrate a clear trade-off between detection and localization. Training with precise temporal labels improves boundary-overlap localization by approximately one-third, while increasing conventional file-level error by around 10 percentage points. These findings highlight that strong utterance-level performance does not necessarily imply accurate localization. SAMVAD-DFprovides both evaluation protocols and diagnostic tools to facilitate rigorous research on the next generation of audio deepfake detection.

1 INTRODUCTION

The proliferation of neural text-to-speech (TTS) and voice conversion (VC) technologies has dramatically lowered the barrier to creating convincing audio deepfakes. As these technologies mature, the threat landscape is shifting from fully synthetic utterances to precise partial manipulations, i.e., recordings in which only a fraction of the audio has been fabricated. This is especially critical in multispeaker settings, where a single malicious speaker's turn can be substituted while the rest of the conversation remains untouched. Such modification, while local and minute, has a long-lasting impact on the spread of misinformation: a synthetic statement can be seamlessly embedded within authentic speech, manipulating facts while preserving the acoustic trust of the surrounding context. The real-world urgency of this threat is already evident in recent electoral disruptions. In September 2023, a targeted audio deepfake was deployed just days before the Slovakian parliamentary elections to manipulate public perception (Meaker, 2023; Aarthi & Agarwal, 2025). In January 2024, an AI-generated robocall weaponized the cloned voice of a U.S. presidential candidate in an attempt to suppress voter turnout during the New Hampshire primary (Swenson, 2024).

Countermeasure (CM) research has responded with a succession of challenge datasets (Wang et al., 2020; Zhang et al., 2022; Yi et al., 2023). Zhang et al. (2022) were the first to formulate partial spoofing as a systematic, scaled benchmark task, building on an earlier, narrower demonstration that partial fakes are detectable at all: Yi et al. (2021) introduced the Half-Truth (HAD) dataset, showing

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.

that replacing a handful of words in an otherwise genuine utterance is already substantially harder to catch than full-utterance synthesis. Subsequent work, such as LAV-DF (Cai et al., 2022) and ADD 2023 (Yi et al., 2023), extended localization benchmarks further, and dedicated localization architectures, TDL (Xie et al., 2024) and BAM (Zhong et al., 2024), were proposed directly on top of PartialSpoof's frame-level protocol. However, this entire line of work is overwhelmingly anchored in single-speaker English speech and controlled studio recordings. This gap is especially pronounced for Hindi, a language spoken by over 600 million people that remains severely underrepresented in anti-spoofing research. This omission is critical given the documented surge in highly targeted political deepfakes, including unauthorized voice clones of public figures, during the 2024 Indian general elections (Landrin, 2024).

To address these gaps, we introduce SAMVAD-DF, the first large-scale partial-deepfake benchmark grounded in the acoustic reality of in-the-wild, multispeaker Hindi conversation. SAMVAD-DF makes five contributions:

1. **Low-resource language coverage**: Native Hindi, including spontaneous Hinglish code-switching;
2. **Multispeaker, in-the-wild audio**: Sourced from YouTube across three diverse domains with up to 76 speakers per recording;
3. **Multi-granularity manipulation**: Three modes spanning full-speaker replacement, partial-segment substitution, and an adversarial intra-segment word splice;
4. **Dual-system annotation verification**: Cross-validated speaker diarization (Pyannote and NeMo TitaNet) on the validation and test splits, isolating a high-purity evaluation track from diarization-confounded regions;
5. **A unified detection and localization evaluation framework:** SAMVAD-DF supports both conventional file-level detection and fine-grained segment/frame-level localization, complemented by diagnostic protocols that disentangle labeling errors from genuine model limitations.

Beyond the dataset, SAMVAD-DF introduces a refined temporal annotation framework for partial deepfakes. Starting from window-level metadata, we further refine manipulated boundaries through human-assisted verification to obtain precise temporal labels for localization. This refined protocol enables more faithful segment- and frame-level evaluation while also providing a complementary file-level benchmark. Together with analyses of splice-induced acoustic characteristics and baseline evaluations under both standard and refined annotations, SAMVAD-DF establishes a rigorous benchmark for localized audio deepfake detection.

2 RELATED WORK

The trajectory of audio deepfake detection has been defined by a transition from the laboratory to the real world. The ASVspoof series (Wang et al., 2020) established the canonical benchmark for the field but evaluated models on clean, single-speaker, fully synthetic utterances. PartialSpoof (Zhang et al., 2022) pioneered the partial-manipulation task at scale, building on the earlier, smaller-scale demonstration of the same threat model in Half-Truth (Yi et al., 2021); both, however, are derived entirely from clean, English-only studio corpora (VCTK and AISHELL-3, respectively) and lack the acoustic unpredictability of real-world speech. Recent initiatives have pushed into more complex environments: ADD 2023 (Yi et al., 2023) expanded localization benchmarks into noisier conditions, and MsCADD (Ahmed et al., 2025) introduced a taxonomy for multispeaker conversational deepfakes, though limited to fully synthetic, two-speaker clips. HQ-MPSD (Li et al., 2025a) and LENS-DF (Liu et al., 2025) represent the current state of the art in rigorous benchmarking, providing a multilingual corpus with linguistically coherent splice points and longer-duration, noisy multi-speaker audio, respectively; neither, however, captures the dense conversational chaos, code-switching, and low-resource realities of South Asian broadcast media that SAMVAD-DF targets. PartialEdit (Zhang et al., 2025) studies partial deepfakes produced by neural speech editing rather than TTS splicing, and similarly finds that detectors trained on clean partial-spoof corpora exploit the artifacts of the underlying neural codec rather than the edit itself—a finding that motivates our attention to compression-domain confound.

On the modeling side, early countermeasures optimized for the global classification paradigm established by ASVspoof. AASIST (Jung et al., 2022) pools temporal features into a single binary score per utterance. Self-supervised front-ends have since dominated the field- SLS-XLSR (Zhang et al., 2024), NACL-SDD (Kim et al., 2025b), and enhanced bonafide modeling (EBM) (Kim et al., 2025a) leverage massive pretrained representations (Babu et al., 2022) to detect subtle acoustic anomalies. But since these architectures compress a sequence into a single global embedding, they are structurally ill-suited to localizing a manipulation boundary. Dedicated localization architectures address this directly: TDL (Xie et al., 2024) learns a frame-level embedding space that separates real and fake frames via temporal convolution, and BAM (Zhong et al., 2024) adds an explicit boundary-detection branch that gates frame-wise attention. Consequently, we also have TDAM Li et al. (2025b) and TRACE Khan et al. (2026), which explicitly model temporal dependencies to isolate intra-utterance splice boundaries.

Shortcut learning, where models exploit unintended data artifacts (container metadata, silence padding, channel compression) rather than semantic content, is well documented in digital forensics (Geirhos et al., 2020; Smeu et al., 2025). We characterize an acoustic footprint of a partial deepfake compilation: the global re-encoding artifact introduced during synthesis, which is addressed by a standard pipeline matching. By quantifying these effects and evaluating them with a paired-window control, SAMVAD-DF provides a diagnostic framework for distinguishing compilation artifacts from genuine localization performance.

3 PROPOSED MULTISPEAKER AUDIO DEEPFAKE DATASET

3.1 REAL DATA COLLECTION AND PROCESSING

To prevent high-capacity networks from exploiting domain-specific acoustic shortcuts such as auditorium reverberation or formal intonation, we deliberately anchor SAMVAD-DF in the acoustic chaos of in-the-wild speech. We source 107.4 hours of audio across 51 manually vetted YouTube videos spanning three ecologically diverse domains: parliamentary debates, panel discussions, and unscripted podcasts. Raw audio is standardized to 16 kHz Pulse-Code Modulation (PCM) with zero codec compression or gain normalization applied at this stage, preserving the natural acoustic container required for downstream evaluation. Transcribing and aligning this unscripted, heavily Hindi-English code-switched ("Hinglish") data is handled with WhisperX (Bain et al., 2023) under targeted bilingual prompting, backed by custom timestamp interpolation to repair alignment dropouts at non-Devanagari token transitions. Initial speaker boundaries are extracted with Pyannote (Bredin, 2023), smoothed with a 5-word sliding-window majority vote to remove transient turn-taking glitches, and split at the smoothed speaker changes; fragments under 1.5 s are merged into neighboring segments.

3.1.1 GUARANTEEING ANNOTATION INTEGRITY: DUAL-SYSTEM VERIFICATION

Overlapping speech pollution in ground-truth speaker boundaries would confound temporal localization metrics directly, so we apply a strict dual-system verification protocol to the validation and test splits (27.68 hours). NeMo's TitaNet-Large Clustering Diarizer (Koluguri et al., 2022) serves as an independent second-opinion system: a Pyannote segment is validated only if a single dominant NeMo speaker covers at least 50% of active speech and secondary speakers contribute under 20%. This yields a bifurcated protocol – the full Pyannote-annotated **Standard Track**, representing noisy, in-the-wild conditions, and a **NeMo-Verified Track** restricted to segments passing the strict threshold.

Across the 11,062 segments in val/test, 10,010 (90.5%) pass verification, retaining 25.25 hours (91.2%) of annotated speech. Of the 1,052 dropped segments, 94.8% fail due to genuine crosstalk, rapid overlapping turns in domains like parliamentary debate that Pyannote smooths over, but NeMo's finer-grained embeddings resolve into distinct co-active speakers. These are not annotation errors; they are exactly the ambiguous regions that would otherwise confound localization metrics, and separating them into their own track gives a mathematically verified subset for pristine evaluation.

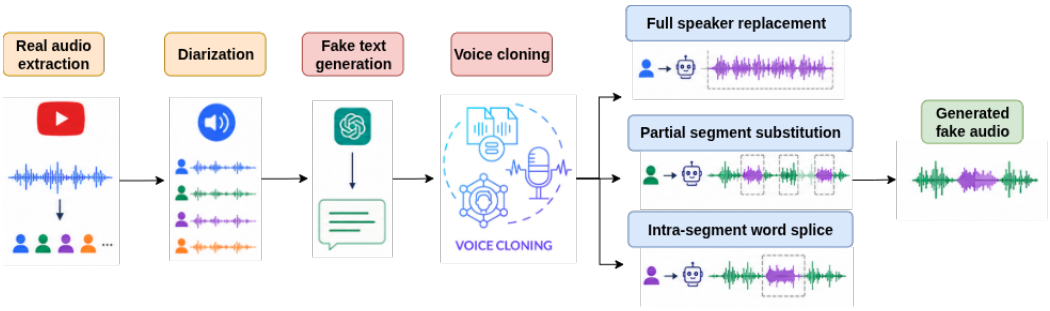

Figure 1: End-to-end pipeline for generating partial and full audio deepfakes. Authentic multi-speaker audio is diarized and conditionally manipulated using LLM-generated transcripts and duration-controlled voice cloning. Manipulations are injected via three distinct splicing modes to produce densely labeled datasets for robust forensic benchmarking.

3.2 Proposed Partial Deepfake Generation

We generate partial deepfakes via a four-stage pipeline: context-aware text generation, high-fidelity voice cloning, spectral envelope matching, and micro-crossfade splicing as illustrated in Figure 1.

Chunking. Source audio is segmented into ∼300-second chunks, with boundaries placed at the mid-point of inter-segment gaps to avoid bisecting active speech; a chunk is retained only if it contains at least two unique speakers.

Fake text generation. For randomly selected target speakers, Qwen2.5-7B-Instruct (Qwen et al., 2025) generates semantically consistent Hindi paraphrases matching the original word count within $\pm 10\%$, aligning synthesized duration with the original utterance and avoiding artifact-inducing time-stretching downstream. Because our synthesis engine mispronounces or skips embedded numerals, we enforce strict script purity (no digits, no residual Latin characters), with temperature annealing (0.3→0.7 over 5 retries) if length or lexical constraints are not met on the first pass.

Voice cloning and duration control. We synthesize with OmniVoice (Zhu et al., 2026), a zero-shot multilingual TTS system, using a single high-fidelity engine deliberately to prevent detectors from relying on low-quality synthesis artifacts as a shortcut. A 3–10 s reference anchor is drawn from the target's own authentic speech in the same chunk. OmniVoice's native duration control dynamically scales synthesis length by the character-ratio between fake and original text, avoiding the phase artifacts introduced by post-hoc time-stretching. All samples were manually verified for phonetic naturalness and temporal alignment.

Spectral matching and splicing. We apply bounded RMS energy matching (± 12 dB) between original and synthesized segments to suppress salient amplitude discontinuities at the boundary, then splice with a 15 ms micro-crossfade – short enough to avoid blending real and fake content. The completed 5-minute chunk is exported via `pydub`, which invokes FFmpeg to re-encode the file; Section 4.3 details why this step matters far more than it looks.

3.3 Multi-Granularity Manipulation Modes

- **Mode 1 – Full-Speaker Replacement.** Every utterance of a target participant is synthesized – the most tractable case.
- **Mode 2 – Partial Segment Substitution.** 1–3 isolated utterances (minimum 3 words) of the target speaker are replaced, simulating a targeted factual manipulation embedded in otherwise authentic speech.
- **Mode 3 – Intra-Segment Word Splice.** The most adversarial tier: 1–2 words, or up to the middle 70% of a segment in higher-variance cases, are substituted, with a minimum 30% word overlap preserved so the original opening and closing audio remains intact. This forces localization to operate without the benefit of natural prosodic pauses.

A single 5-minute bona fide anchor chunk is multiplexed across all three modes (and potentially multiple target speakers), yielding 3,366 perfectly matched pairs (6,732 chunks) from the original 107.4 hours.

3.4 Dataset Statistics and Splits

Table 1: SAMVAD-DF split overview. The Verified Track is the high-confidence NeMo subset available exclusively for validation and test.

Split	Total	Bona Fide	Spoof (Standard)	Spoof (Verified)
Train	4,458	2,229	2,229	–
Val	1,260	630	630	252
Test	1,014	507	507	168
Total	**6,732**	**3,366**	**3,366**	**420**

Evaluation-pool clarification. On the Verified Track, the 168 (test) / 252 (val) NeMo-audited spoof chunks are scored against the *full, unfiltered* bona fide pool of 507 (test) / 630 (val) files – i.e., the Verified Track filters which spoof chunks are eligible for the high-purity localization claim, but deliberately does not additionally filter bona fide files, since bona fide files carry no diarization-dependent manipulation label to verify.

The three source domains—parliamentary debate (15–76 speakers/recording, heavy reverberation and interruption), panel discussion (3–12 speakers, structured but overlapping cross-talk), and podcast/interview (2–4 speakers, clean near-field acoustics) – are partitioned by source recording (not by utterance) to guarantee generalization to unseen acoustic environments, which induces an intentional domain shift toward panel discussions in the evaluation splits (Table 2). Each of the three attack types contributes exactly one-third of spoof chunks in every split, and this 1:1:1 balance is preserved within the Verified Track.

Table 2: Chunk distribution by domain and split. Bona fide and spoof counts are identical (1:1 pairing) in every cell.

Domain	Train	Val	Test	Total
Commentary / Interview / Podcast	942	321	216	1,479
Panel Discussion	552	303	261	1,116
Parliament Debate	735	6	30	771
Total	**2,229**	**630**	**507**	**3,366**

4 Benchmark Experiments and Diagnosis

We evaluate seven countermeasure architectures spanning global SSL classifiers (EBM (Kim et al., 2025a), XLSR-Mamba (Xiao & Das, 2025), SLS-XLSR (Zhang et al., 2024), NACL-SDD (Kim et al., 2025b)), a zero-shot anomaly detector (TRACE (Khan et al., 2026)), and temporal/boundary-aware localization models (AASIST (Jung et al., 2022), TDAM (Li et al., 2025b), and BAM (Zhong et al., 2024)). All supervised baselines are executed via each method's original, publicly released codebase and default training recipe against the SAMVAD-DF training split, with no architectural modification – we report every baseline's original inherited hyperparameters (Appendix A). All evaluation uses a unified 4-second sliding window with a 1-second stride (Transformer/SSL backbones cannot ingest 5-minute files directly), aggregated to a file-level decision.

4.1 Refined Temporal Metadata for Localized Deepfake Detection

Precise temporal supervision is essential for partial deepfake localization. Samvad-DF therefore provides two complementary annotation protocols. The **Standard Track** uses window-level labels derived automatically from manipulation metadata, while the **Refined Track** further improves these

temporal boundaries through human-assisted verification, yielding higher-fidelity localization annotations without altering the underlying audio. To quantify the impact of refined temporal metadata, we retrain AASIST using the refined window labels while keeping the architecture and optimization unchanged. On the Samvad-DF test set, the refined model attains 45.70% EER on the Standard Track and 47.62% EER / 0.5493 AUC on the NeMo-Verified Track, compared with 26.43% / 28.40% under the original metadata protocol (Table 3).

The refined annotations reveal a complementary gain in localization. We infer from Table 4, global Seg-AUC increases from 0.7978 to 0.8958, while boundary-overlap mIoU improves from 0.2233 to 0.2337 (approximately 38% relative improvement). These results demonstrate that higher-quality temporal supervision substantially improves boundary localization, even though the corresponding file-level detection task becomes more challenging. Rather than replacing the original protocol, SAMVAD-DF reports results under both annotation tracks. Together with segment- and frame-level localization metrics, this dual-protocol evaluation enables a more comprehensive assessment of partial audio deepfake detectors by jointly measuring utterance-level detection and temporal localization.

4.2 AGGREGATION STRATEGY

We empirically compared Max, Mean, and 95th-percentile pooling for aggregating window scores to a file-level decision. Max-spoof pooling is the strongest choice for this task (26.4% EER vs. 42.8% for Mean, 36.5% for 95th-percentile), because it isolates a sparse, diluted deepfake signal within a broad 5-minute context rather than averaging it away. This choice is not free, however: because each 5-minute file is scored across $\sim$297 overlapping windows, max-pooling is structurally asymmetric, a single spurious window anywhere in an authentic file is enough to flip the file-level decision. We quantify this risk directly rather than leaving it implicit. The global per-window false-positive rate at the operating threshold is 2.32%, but these false alarms are not independently distributed: only 26% of bona fide files contain any false-alarm window, but the files that do average 28 false-alarm windows each (up to 184), concentrated in specific difficult domains. Max-pooling remains the empirically best aggregator despite this clustering, but the clustering itself is a first-order fact about the evaluation protocol that any EER number derived from it should be read alongside.

4.3 PREVENTING THE CODEC SHORTCUT, AND QUANTIFYING WHAT REMAINS

In early pipeline iterations, high-capacity global classifiers achieved artificially low EER (0.00% on raw, unmitigated exports, and still $\sim$15% after an intermediate padding fix) while exhibiting near-random localization. We traced this to a *global re-encoding bias*: because our generation pipeline invokes FFmpeg to embed the fake audio into the container, spoof files carried a distinct codec signature that classifiers learned to detect in place of the actual manipulation. We close this with strict pipeline matching; every bona fide chunk passes through the identical FFmpeg resampling, chunking, and alignment pipeline as spoof chunks. Stripped of this shortcut, the top-performing baselines bottleneck at $\sim$26% file-level EER (Table 3).

Further footprints are intrinsic to the act of splicing zero-shot synthetic speech into a real, previously compressed, reverberant recording.

Ambient dropout. Because our pipeline applies bounded RMS matching but does not perform RIR convolution or ambient-noise infilling at the insertion point, the ambient noise floor measurably drops during a splice: we measure a 12.44 dB dropout inside the spliced region relative to the paired bona fide anchor. To check whether this alone constitutes a trivial shortcut, we trained a single-feature classifier on this dropout measurement in isolation; it reaches only 0.5785 AUC, indicating the dropout is largely masked by natural speech-energy fluctuation and is not, on its own, a reliable detection cue, though we regard this as an initial probe rather than an exhaustive acoustic audit. Injecting pristine synthetic speech into a degraded, reverberant acoustic container will introduce localized artifacts of this kind in *any* partial-deepfake benchmark built this way; the contribution here is not that SAMVAD-DF is artifact-free, but that it is the first, to our knowledge, to measure exactly where the boundary between genuine phonetic detection and residual acoustic bookkeeping sits, so that future work can evaluate around it rather than through it.

Table 3: Global test-set performance. The Standard Track evaluates the full Pyannote-annotated test set. The Verified Track scores the 168 high-confidence segments.

Model	Standard Track		Verified Track	
	EER (%)	AUC	EER (%)	AUC
Original window-labeling convention				
SLS-XLSR	26.33	0.816	27.38	0.801
AASIST	26.43	0.804	28.40	0.781
EBM	33.92	0.739	31.95	0.749
XLSR-Mamba	48.22	0.516	47.63	0.503
NACL-SDD	48.99	0.530	50.00	0.509
Unaffected by window-level labeling (zero-shot)				
TRACE	47.93	0.511	47.34	0.525
Refined window labeling				
TDAM	26.82	0.820	24.40	0.825
AASIST (refined labels)	45.70	0.583	47.62	0.549
BAM (refined labels)	55.62	0.527	55.62	0.529

4.4 Genuine Localization: Segment- and Frame-Level Metrics

File-level EER answers "is this file spoofed," not "where." To evaluate localization directly, we compute Seg-AUC and mIoU (Jaccard overlap between predicted and ground-truth manipulated spans) under a 2.0-second window, 0.5-second hop protocol against exact JSON manipulation boundaries, for every model whose output admits a continuous per-window score.

Table 4: Localization performance. Seg-AUC and mIoU are frame-level discriminability and boundary-overlap metrics, computed independently of the file-level classification protocol in Table 3. The two AASIST rows share one architecture and differ only in which labeling convention (Section 4.1) trained them.

Model	Architecture Type	Global Seg-AUC	Global mIoU	Standard Seg-AUC	Verified Seg-AUC
AASIST (original labels)	Graph / GAT	0.7978	0.2233	0.8261	0.7150
AASIST (corrected labels)	Graph / GAT	0.8958	0.2337	0.9417	0.7584
TDAM	Temporal Attention	0.7706	0.2061	0.7983	0.6894
EBM	Wav2Vec2 + Linear	0.7683	0.1976	0.7965	0.6815
TRACE	Zero-Shot Foundational	0.5151	0.0683	0.5151	0.5152
NACL-SDD	XLSR + Conformer	0.3651	0.0008	0.3524	0.4035
SLS-XLSR	Wav2Vec2 + ResNet	0.2236	0.0005	0.1977	0.2996
BAM	Boundary-Supervised	0.9029	0.0047	0.8941	–

The AASIST rows are, we think, the most consequential single result in the paper: identical architecture, identical test set, and the only variable across rows is which labeling convention trained the model. Under the original convention, AASIST reaches a respectable Seg-AUC of 0.7978 but a modest boundary-overlap mIoU of 0.2233, it can tell that a file is probably touched, without precisely bounding where. Correcting the training labels raises Global Seg-AUC to 0.8958 (0.9417 on the Standard Track alone) and mIoU by roughly a third, even though the same correction *worsens* file-level EER (Table 3). That is exactly the trade-off Section 4.1 describes: the original labels reward a coarse, container-level cue that inflates file detection at the cost of genuine localization. **TDAM and EBM localize competently under the original protocol as well** (Seg-AUC 0.77 each), consistent with architectures whose pooling is shallow enough to preserve frame-level resolution even without a corrected-label retrain. SLS-XLSR and NACL-SDD, by contrast, are near-random or worse at localization (0.22 and 0.37 Seg-AUC, mIoU $\approx$0) despite reporting file-level EER competitive with AASIST's original-convention number in Table 3. TRACE, evaluated fully zero-shot, sits at chance (0.5151).

4.5 DIAGNOSING THE WORD-SPLICE COLLAPSE

Full-speaker replacement is the most tractable manipulation mode; SLS and AASIST reach 7.10% and 11.44% EER respectively (Table 5). Word splices collapse this to near-chance across every model (SLS: 48.82%). Two distinct explanations are consistent with this collapse: genuine synthesis quality is too subtle to detect at 1-2 word granularity, or the sliding-window pooling protocol itself is accumulating false alarms that swamp a real but faint signal.

We designed a true-positive argmax stratification for exactly this purpose, applied to the original-convention AASIST checkpoint: for every file the model correctly flags as spoof, we check whether the single highest-scoring window (the one that determines the max-pooled file-level decision) actually overlaps the ground-truth manipulated region. For partial-segment attacks, 94.1% of files were correctly flagged, and among those, the argmax window overlapped the true fake region 96.2% of the time (against an 8.3% null from random placement), clear evidence of genuine localization. For word-splice attacks, only 28.4% of files were correctly flagged at all, and among those, the argmax window overlapped the fake region just 14.6% of the time (against a 4.7% null), barely above chance. As a targeted case study, we further computed a strict window-level AUC restricted to windows with $\geq$50% true overlap: 0.8691 for partial segments versus 0.5344 (chance) for word splices, corroborating the file-level pattern at the window level directly.

This rules out false-alarm accumulation as the primary driver: the model is not failing because pooling drowns a real signal in noise; when it does flag a word-splice file correctly, it is very largely not looking at the manipulated region at all. Read alongside Section 4.1, this checkpoint's behavior is exactly what the original labeling convention would predict, rewarded for reacting anywhere in a spoof file's container rather than at the manipulation itself, it did. Whether the corrected-label checkpoint's markedly better localization (Table 4) closes this specific gap on word splices, the hardest and sparsest attack tier, is the direct next step in this analysis, and we report the rerun once the corrected-label model's per-attack-type breakdown is available.

Table 5: Model degradation by attack granularity (Standard Track EER, %). SLS-XLSR and AASIST figures reflect the original window-labeling convention (Section 4.1); a per-attack-type breakdown for the corrected-label AASIST checkpoint is not yet available and is reported once complete.

Attack Type	**SLS-XLSR**	**TDAM**	**AASIST (orig. labels)**	**BAM**
Full Speaker	7.10	11.24	11.44	47.76
Partial Segments	13.02	18.93	14.79	48.21
Word Splice	48.82	42.01	48.52	49.01

4.6 THE LEXICAL SIGNATURE CONFOUND

Because our LLM generation pipeline is constrained to pure Devanagari output (to avoid TTS hallucination) while authentic SAMVAD-DF audio contains heavy Hinglish code-switching, a text-level shortcut is possible: a TF-IDF logistic regression classifier on ground-truth transcripts alone reaches 93.98% AUC, trivially exploiting the presence of English tokens as an authenticity signal. We control for this with a domain-restricted ablation: isolating evaluation to Panel Discussions (87.4% code-switching) versus Podcasts (58.3%) leaves TDAM's EER essentially unchanged (26.13%$\rightarrow$24.00%), indicating the acoustic models are not exploiting the lexical artifact and are evaluating genuine acoustic structure.

4.7 CROSS-DATASET AND CROSS-GENERATOR GENERALIZATION

Evaluating SAMVAD-DF-trained baselines zero-shot on LENS-DF (Liu et al., 2025) (17 distinct English TTS systems, no fine-tuning) shows substantial degradation. TDAM, our clean, label-correction-independent reference point, drops from an in-domain 26.82% to 39.30% EER. AASIST's original-convention checkpoint drops from 26.43% to 37.70%; its corrected-label checkpoint transfers to 36.90% EER / 0.6716 AUC / 0.6717 average precision—essentially the same cross-dataset transfer performance as the original-convention checkpoint, despite scoring nearly twenty points worse in-domain (Table 3). We read this as further evidence that the original convention's in-domain advantage is specific to the labeling artifact itself, rather than a generalizable detection

capability: it does not survive the move to a different dataset, even when it looks dominant at home. Conversely, on SAMVAD-DF's own cross-generator (XTTS) protocol, same language and acoustic environment, different synthesizer, SLS-XLSR achieves 9.00% EER, suggesting its layer-wise attention over 24 XLSR layers tracks vocoder-agnostic phase discontinuities rather than overfitting to OmniVoice's specific spectral fingerprint (an original-convention number; we plan to revisit it once SLS-XLSR's corrected-label retrain completes). The gap between near-in-distribution generator transfer and cross-lingual/cross-domain transfer underscores why a large, acoustically diverse, non-English anchor benchmark is useful independent of any single model's performance on it.

5 LIMITATIONS AND FUTURE WORK

We deliberately used one high-fidelity TTS engine (OmniVoice) to prevent detectors from relying on low-quality synthesis artifacts as a shortcut; future work extending to multiple generators (XTTS (Casanova et al., 2024), VALL-E (Han et al., 2024)) would test cross-generator robustness without that confound.

Broader impact. The weaponization of generative audio threatens democratic processes, courtroom evidence, and institutional trust, and this threat is asymmetric: low-resource language communities are especially vulnerable due to the scarcity of detection tools calibrated to their linguistic and acoustic realities. By targeting Hindi – a language spoken by over 600 million people that experienced a documented surge in political voice cloning during the 2024 Indian general elections (Landrin, 2024) – SAMVAD-DF addresses a concrete geographic disparity in the field's coverage.

6 CONCLUSION

As malicious actors pivot from fully synthetic utterances to precise, intra-segment manipulations, standard deepfake-detection paradigms and standard deepfake-*evaluation* paradigms are both becoming inadequate. SAMVAD-DF contributes a large-scale, in-the-wild, multispeaker Hindi partial-deepfake benchmark, and, we would argue, an equally important methodological finding: the window-labeling convention this task inherited from fully-synthetic benchmarks inflates apparent file-level detection while actively suppressing genuine localization, and the two effects are large enough, and move in opposite enough directions to change which architecture looks best depending on which convention evaluates it. We show this directly on a single architecture, holding everything but the labels fixed, and find that current state-of-the-art models' apparent detection of the hardest attack tier, sub-second word splices, is, to a large extent, not detection of the splice at all, but detection of the very shortcut the corrected labels remove.

ETHICS STATEMENT

SAMVAD-DF is built entirely from publicly available YouTube broadcast content. Before finalizing the source list, we conducted a source-level rights assessment of every channel and video included, and we provide an opt-out mechanism through which any speaker or channel owner can request that their audio be excluded from future distributions of the dataset. Because the dataset's purpose is to train and evaluate *detectors*, we do not release a runnable deepfake-generation pipeline; the OmniVoice-based synthesis procedure is described in Section 3 for scientific transparency and reproducibility, but the code to run it end-to-end is not distributed. Access to SAMVAD-DF's code and data is gated behind a research-use agreement restricting use to non-commercial detection research and prohibiting redistribution or use in generating deceptive or misleading content.

REPRODUCIBILITY STATEMENT

Every result in this paper is reconstructable from Section 4 and Appendix A: Section 4 fixes the shared evaluation protocol (4-second windows, 1-second stride, max-spoof pooling to a file-level decision over the 1,014-file test set split into Standard and NeMo-Verified Tracks) and, independently, the segment- and frame-level localization protocol (2.0-second window, 0.5-second hop against JSON manipulation boundaries) used for the Seg-AUC/mIoU results; Section 4.1 states

the exact criterion for the corrected window labels (a window is spoof iff it has nonzero overlap with a JSON manipulation span), sufficient to reimplement the correction independent of our rebuild script; and Appendix A lists, per baseline, whether it consumes the shared window-level TSV or a native/independent loader, confirms each uses its original public architecture, loss, and optimizer unmodified with only SAMVAD-DF substituted as input, and states the compute environment (2× NVIDIA H200 NVL GPUs, 141 GB VRAM each). Data and code access is governed by the research-use agreement described in the Ethics Statement.

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 *International Joint Conference on Biometrics*, pp. 1–11. IEEE, 2025.

Alabi Ahmed, Vandana Janeja, and Sanjay Purushotham. Multi-speaker conversational audio deepfake: Taxonomy, dataset and pilot study. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1373–1378. IEEE, 2025.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech*, pp. 2278–2282, 2022.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *Interspeech*, 2023.

Hervé Bredin. pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech*, pp. 1983–1987. ISCA, 2023.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–10. IEEE, 2022.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. Xtts: a massively multilingual zero-shot text-to-speech model. In *Proc. Interspeech*, pp. 4978–4982, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Bing Han, Long Zhou, Shujie LIU, Sanyuan Chen, Lingwei Meng, Yanmin Qian, Eric Liu, sheng zhao, Jinyu Li, and Furu Wei. VALL-e r: Robust and efficient zero-shot text-to-speech synthesis via monotonic alignment. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024. URL https://openreview.net/forum?id=xvORqaYDgL.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *IEEE international conference on acoustics, speech and signal processing*, pp. 6367–6371. IEEE, 2022.

Awais Khan, Muhammad Umar Farooq, Kutub Uddin, and Khalid Malik. Trace: Training-free partial audio deepfake detection via embedding trajectory analysis of speech foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7405–7414, 2026.

Seung-bin Kim, Hyun-seo Shin, Jungwoo Heo, Chan-yeong Lim, Kyo-Won Koo, Jisoo Son, Sanghyun Hong, Souhwan Jung, and Ha-Jin Yu. Enhancing audio deepfake detection by improving representation similarity of bonafide speech. In *Interspeech*, pp. 2250–2254, 2025a.

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, 2025b.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In *IEEE international conference on acoustics, speech and signal processing*, pp. 8102–8106. IEEE, 2022.

Sophie Landrin. India's general election is being impacted by deepfakes. *Reuters*, 2024. URL https://www.lemonde.fr/en/pixels/article/2024/05/21/india-s-general-election-is-being-impacted-by-deepfakes_6672168_13.html#. Accessed: 2026-05-03.

Menglu Li, Majd Alber, Ramtin Asgarianamiri, Lian Zhao, and Xiao-Ping Zhang. Hq-mpsd: A multilingual artifact-controlled benchmark for partial deepfake speech detection. *arXiv preprint arXiv:2512.13012*, 2025a.

Menglu Li, Xiao-Ping Zhang, and Lian Zhao. Frame-level temporal difference learning for partial deepfake speech detection. *Signal Processing Letters*, 2025b.

Xuechen Liu, Wanying Ge, Xin Wang, and Junichi Yamagishi. Lens-df: Deepfake detection and temporal localization for long-form noisy speech. *arXiv preprint arXiv:2507.16220*, 2025.

Morgan Meaker. Slovakia's election deepfakes show AI is a danger to democracy. *Wired*, 2023. URL https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/. Accessed: 2024-05-01.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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*, 2025.

Ali Swenson. Fake Joe Biden robocall tells New Hampshire voters not to vote tuesday. *The Associated Press*, 2024. URL https://edition.cnn.com/2024/01/22/politics/fake-joe-biden-robocall. Accessed: 2024-05-01.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020.

Yang Xiao and Rohan Kumar Das. Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection. *Signal Processing Letters*, 32:1276–1280, 2025.

Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. An efficient temporary deepfake location approach based embeddings for partially spoofed audio detection. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 966–970. IEEE, 2024.

Jiangyan Yi, Ye Bai, Jianhua Tao, Haoxin Ma, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu. Half-Truth: A Partially Fake Audio Detection Dataset. In *Interspeech*, pp. 1654–1658, 2021. doi: 10.21437/Interspeech.2021-930.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, et al. Add 2023: the second audio deepfake detection challenge. In *CEUR Workshop Proceedings*, volume 3597, pp. 125–130, 2023.

Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi. The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:813–825, 2022.

Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised xls-r and sls classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6765–6773, 2024.

You Zhang, Baotong Tian, Lin Zhang, and Zhiyao Duan. PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing . In *Interspeech*, pp. 5353–5357, 2025. doi: 10.21437/Interspeech.2025-942.

Jiafeng Zhong, Bin Li, and Jiangyan Yi. Enhancing Partially Spoofed Audio Localization with Boundary-aware Attention Mechanism. In *Interspeech 2024*, pp. 4838–4842, 2024. doi: 10.21437/Interspeech.2024-587.

Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. Omnivoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688*, 2026.

A Benchmark Training and Evaluation Details

All supervised baselines are trained on the SAMVAD-DF training split and validated on the corresponding validation split, with input audio standardized to 16 kHz mono PCM, and all models are evaluated on the identical held-out test set (1,014 files: 507 spoof, 507 bona fide). Each baseline retains its architecture, loss, and optimizer exactly as specified in its original public release, with no changes beyond swapping in SAMVAD-DF's data. Data supply differs by baseline, which is the source of the labeling finding in Section 4.1: AASIST (Jung et al., 2022), SLS-XLSR (Zhang et al., 2024), NACL-SDD (Kim et al., 2025b), EBM (Kim et al., 2025a), and XLSR- Mamba (Xiao & Das, 2025) consume pre-built window-level TSVs under the unified 4-second sliding window / 1-second stride / max-spoof pooling protocol; TDAM (Li et al., 2025b) uses a native segment-level loader that slices audio dynamically against exact JSON manipulation boundaries rather than a window TSV; TRACE (Khan et al., 2026) is evaluated fully zero-shot with no SAMVAD-DF-specific training data at all; and BAM's (Zhong et al., 2024) frame-level supervision is precomputed directly from the JSON spans. All experiments ran on an internal cluster with 2× NVIDIA H200 NVL GPUs (141 GB VRAM each).