# CHAINBENCH-ADD: A DELIVERY-AWARE DATASET AND BENCHMARK FOR AUDIO DEEPFAKE DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Existing audio deepfake datasets have greatly expanded evaluation across generators, languages, and domains, but most remain generation-centric and provide limited support for studying post-generation delivery. In practical misuse scenarios, forged audio may undergo platform re-encoding, telephony transmission, or replay-like recapture before detection. Existing delivery-related studies usually model such factors as isolated distortions and often lack control over transcript and speaker conditions. To address this gap, we introduce ChainBench-ADD, a delivery-aware dataset and benchmark for audio deepfake detection. It provides a controlled and reproducible approximation of common delivery factors, including transcoding/re-encoding, communication-channel degradation, playback/re-recording effects, and multi-hop delivery. These factors are organized into five delivery families using reusable operators, ordered templates, and realized chains, while each delivered sample is linked to a clean bona fide or spoof parent under fixed transcript and speaker conditions. The current release contains 941,201 waveforms derived from 55,813 parents and supports five protocol-aware tasks. Experiments show that ChainBench-ADD reveals family-level difficulty differences, sensitivity to controlled local edits, and robustness decay along accumulated delivery paths that conventional pooled evaluation cannot capture. We release the dataset and code to support reproducible delivery-aware evaluation.

## 1 INTRODUCTION

Recent advances in text-to-speech, voice cloning, and speech generation have made audio deepfakes increasingly realistic and accessible, turning audio deepfake detection into an important problem in multimedia security. In response, the community has developed a broad range of benchmarks and datasets that substantially improve evaluation across generators, tasks, languages, and domains Liu et al. (2023); Yi et al. (2024); Frank & Schönherr (2021); Khalid et al. (2021); Müller et al. (2022); Zhang et al. (2021); Sun et al. (2023); Müller et al. (2024); Ma et al. (2024); Wang et al. (2026); Xie et al. (2025); Bhagtani et al. (2025); Ciobanu et al. (2026); Wang et al. (2025a). These resources have considerably strengthened evaluation on the generation side of audio deepfakes.

However, realistic misuse rarely ends at synthesis. In practice, a forged clip may be uploaded to an online platform, transmitted through a telephony channel, or played aloud and re-recorded before it reaches a listener or an automated system. The audio encountered at deployment is therefore often not pristine generated speech, but speech after post-generation delivery. This creates a mismatch between common benchmark conditions and realistic use scenarios. Some recent studies have started to consider communication effects, codec distortions, or replay-like conditions Shi et al. (2025); Müller et al. (2025); Delgado et al. (2025), but they mostly examine delivery through isolated effects. While useful, such settings do not fully reflect the fact that post-generation delivery is often a structured process in which multiple operations are combined, parameterized, and ordered. Platform re-encoding, telephony transmission, and replay-like recapture may affect forensic cues differently, and their impact may depend on the particular delivery route.

These considerations suggest the need for a benchmark that treats post-generation delivery as a structured evaluation dimension: one that spans multiple common delivery families through controlled, reproducible approximations, represents delivery as ordered chains rather than isolated perturbations, and supports controlled analysis under matched conditions. Such control is important



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.

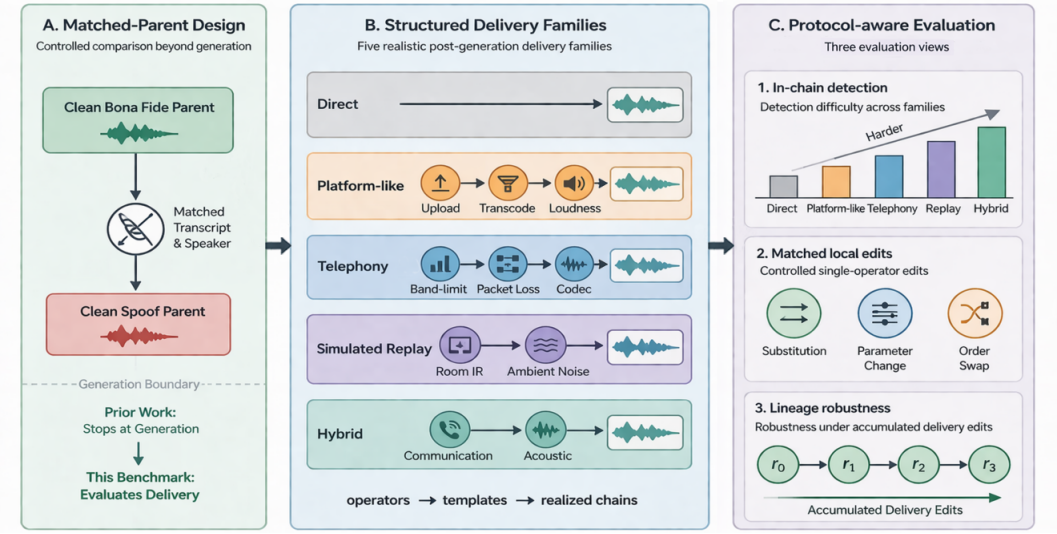



Figure 1: ChainBench-ADD models post-generation delivery as ordered chains under matched-parent control, going beyond generation-centric benchmarks and isolated delivery effects.

because, without matched clean parents, delivery effects can be entangled with differences in linguistic content, speaker identity, or source audio quality, making it difficult to attribute detector behavior specifically to delivery.

To this end, we introduce ChainBench-ADD, a delivery-aware dataset and benchmark for audio deepfake detection. It provides a controlled and reproducible approximation of post-generation delivery through reusable operators, ordered templates, and realized chains across five delivery families: direct, platform-like, telephony, simulated replay, and hybrid. Each delivered sample remains linked to a clean bona fide or spoof parent under matched-parent control, enabling controlled comparison across delivery scenarios. Figure 1 summarizes the benchmark design.

The current release contains over 941k waveforms derived from 55,813 parents. From the shared metadata, we define five evaluation tasks: in-chain detection, three matched local interventions (operator substitution, parameter perturbation, and order swap), and lineage-based delivery robustness. A leave-one-template-out protocol further tests transfer to unseen templates within a family. Together, these components support controlled, deployment-motivated delivery-aware evaluation for audio deepfake detection.

The main contributions of this work are as follows:

- We introduce ChainBench-ADD, a large-scale delivery-aware dataset and benchmark that represents common post-generation factors through a controlled, structured approximation spanning five delivery families, reusable operators, ordered templates, and realized chains.
- We design a matched-parent, metadata-driven benchmark framework that isolates delivery effects under fixed transcript and speaker conditions and supports controlled, protocol-aware evaluation within a unified metadata schema.
- We release a protocol-aware resource with explicit lineage annotations, benchmark statistics, perceptual-quality and matched-parent consistency validation, and show through baseline studies that it exposes family-level, local, and path-wise robustness patterns that conventional pooled evaluation cannot capture.

## 2 Related Work

Recent audio deepfake resources have developed along four complementary directions: challenge benchmarks, diversity-oriented datasets, delivery-oriented robustness studies, and reproducibility

Table 1: Comparison of representative audio deepfake benchmarks. ✓ denotes explicit support, (✓) partial support, — absence, and N/A not applicable because the work is primarily an evaluation framework.

| Resource | Focus | Post-generation delivery | Ordered comp. | Matched-parent control |
|---|---|---|---|---|
| ASVspoof 2021 | challenge benchmark | codec/channel/room effects in challenge tracks | — | — |
| ADD 2023 | task benchmark | not delivery-centric | — | — |
| CFAD | generator/domain range | noise + codec condition sets | — | — |
| CodecFake | ALM/codec coverage | generator-side codec effects | — | — |
| ADD-C | comms robustness | codec + packet-loss conditions | (✓) | — |
| ReplayDF | replay robustness | playback + re-recording pipeline | (✓) | (✓) |
| AUDDT | evaluation toolkit | N/A | N/A | N/A |
| Spoof-SUPERB | SSL benchmark | N/A | N/A | N/A |
| **ChainBench-ADD (ours)** | **structured delivery** | **multi-family post-generation delivery** | ✓ | ✓ |

frameworks. Table 1 focuses on the aspects most relevant to this work: delivery scope, structured composition, and matched-parent control.

Early evaluation standards were shaped largely by the ASVspoof series. ASVspoof 2021 Liu et al. (2023) formalized logical-access, physical-access, and deepfake evaluation while incorporating compression, transmission, and acoustic propagation effects; ASVspoof 5 Wang et al. (2026) further broadened speaker, acoustic, and attack diversity. In parallel, ADD 2023 Yi et al. (2024) expanded evaluation beyond utterance-level bona fide/spoof discrimination to include manipulation localization and algorithm recognition, and PartialSpoof Zhang et al. (2021) focused on partially manipulated utterances. Earlier resources such as WaveFake Frank & Schönherr (2021) and FakeAVCeleb Khalid et al. (2021) also helped establish benchmarked evaluation for synthetic and multimodal deepfakes.

A second line of work emphasizes broader generator, language, and domain coverage. In-the-Wild Müller et al. (2022) highlighted the gap between laboratory benchmarks and real online media. MLAAD Müller et al. (2024) expanded multilingual and multi-generator coverage, LibriSeVoc Sun et al. (2023) emphasized vocoder artifacts and self-vocoding conditions, CFAD Ma et al. (2024) introduced a Chinese benchmark under noisy and transcoding conditions, and CodecFake Xie et al. (2025) focused on codec-based and neural-codec generation. More recent datasets Huang et al. (2025); Mawalim et al. (2025); Ciobanu et al. (2026); Wang et al. (2025a) continue this trend toward larger-scale multilingual, cross-domain, and open-world evaluation. Despite that, the main axis of variation in these resources remains the generator, language, or source domain.

More closely related to our setting are delivery-oriented robustness studies. ADD-C Shi et al. (2025) shows that codec compression, packet loss, and communication-channel effects can noticeably affect detector performance in VoIP- and telephony-like scenarios, while ReplayDF Müller et al. (2025) reports similar degradation under playback and re-recording. Related discussion in Delgado et al. (2025) further points to a realism gap between raw generated audio and audio after practical presentation through communication channels or platform pipelines. These studies establish delivery as an important source of performance variation, but they typically examine communication, replay, or platform effects separately rather than as parts of one structured delivery process.

Recent frameworks have also improved reproducibility at the system level. AUDDT Zhu et al. (2025) standardizes cross-dataset evaluation across public corpora, and Spoof-SUPERB Ali et al. (2026) provides a SUPERB-style comparison of self-supervised speech representations for audio deepfake detection. These efforts improve fair and reproducible comparison, but they primarily reorganize or evaluate existing resources rather than introduce a matched-parent benchmark centered on post-generation delivery.

ChainBench-ADD is designed to complement this landscape by modeling post-generation delivery as an ordered process under matched-parent control. To the best of our knowledge, existing representative resources have not jointly provided multiple post-generation delivery families, explicit ordered templates, and matched-parent local analysis within one unified benchmark.

Table 2: Delivery operators, representative parameter settings, and family coverage in ChainBench-ADD.

| Operator | Representative settings | Fam. |
|---|---|---|
| Resample | 16→8, 16→24, 16→8→16, 16→24→16, 16→32→16 kHz | P/T/R/H |
| Band-limit | Narrowband, wideband | T/H |
| Codec | AAC (24/32/48 kbps), Opus (16/24/32 kbps), GSM, $\mu$-law PCM | P/T/H |
| Re-encode | Same/cross-codec AAC/Opus recompression (24/32 kbps) | P/R/H |
| Packet loss | 1/3/5/10% loss; burst 2/3/5; repeat-fade, interpolation, or noise-fill concealment | T/H |
| Noise | White, pink, brown, babble, hiss, hum; 30/20/15/10 dB SNR | R |
| RIR | Small/medium/large rooms; RT60 0.2/0.4/0.6/0.8 s; source–microphone distance 0.5/1/2/3 m | R/H |
| Call-path | Joint channel filtering, codec, packet loss, jitter (0/8/16 ms), and AGC | T |

**Fam.** P = platform-like, T = telephony, R = simulated replay, H = hybrid.

# 3 THE CHAINBENCH-ADD BENCHMARK

## 3.1 OVERVIEW AND TERMINOLOGY

ChainBench-ADD organizes post-generation delivery into families, templates, variants, and realized chains. This hierarchy supports scenario-, route-, local-, and path-level analysis under matched-parent control.

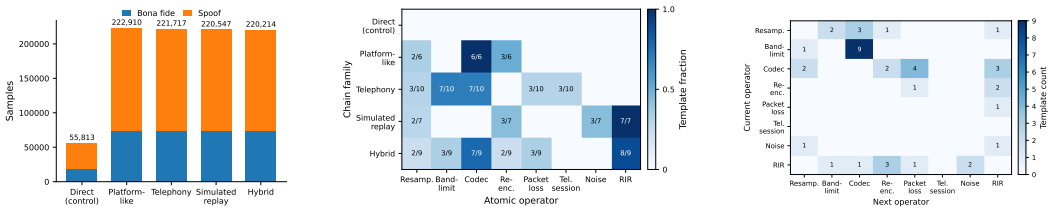
Throughout this section, a delivery operator denotes an atomic or composite transform such as codec compression, re-encoding, packet loss, room reverberation, or additive noise. A template is an ordered operator blueprint within one delivery family, while a variant is one parameterized instantiation. A delivery chain is the realized ordered path that is applied to one parent sample. A delivery family is the top-level scenario category that groups related templates, including direct, platform-like, telephony, simulated replay, and hybrid delivery.

## 3.2 DELIVERY FAMILIES AND CHAIN REALIZATION

ChainBench-ADD is a controlled and reproducible approximation of post-generation delivery rather than an exact replica of any particular platform, network, room, or device. It covers transcoding/re-encoding, communication-channel degradation, playback/re-recording effects, and multi-hop combinations. These factors are organized into five delivery families. Direct is the identity control; platform-like approximates online redistribution and recompression; telephony models bandwidth-limited call transmission with telephony codecs, packet impairment, and compact end-to-end call simulation; simulated replay approximates playback/re-recording in physical spaces through room response and environmental noise; and hybrid combines communication and acoustic effects within a single path. These families are scenario abstractions that may share operators but combine them in different orders and parameter regimes.

These families are built from eight reusable delivery operators: resampling, band-limiting, codec compression, re-encoding, packet loss, additive noise, room impulse response (RIR), and call-path simulation. Many operators are shared across multiple families. For example, codec-like effects appear in both platform-like and telephony delivery, while RIR appears in both simulated replay and hybrid delivery. A template specifies an ordered operator blueprint within one delivery family, a variant instantiates that blueprint with concrete parameter choices, and a realized chain is the waveform-domain rendering of that ordered path for one parent sample. Table 2 summarizes the operator inventory and representative parameter ranges of each operator. For example, a hybrid chain $\mathrm{codec} \rightarrow \mathrm{packet_loss} \rightarrow \mathrm{RIR}$ combines communication and acoustic effects. Within simulated replay, $\mathrm{RIR} \rightarrow \mathrm{noise}$ and $\mathrm{noise} \rightarrow \mathrm{RIR}$ form an order-swap comparison when rendered from the same parent with matched parameters: the operator inventory is unchanged, but its order differs. The full template inventory and parameter pools are given in Appendix A.

All operators are waveform-domain transforms rather than symbolic tags. Codec and re-encode effects use encode–decode round trips, packet loss includes burstiness and concealment, RIR is ap-




(a) Delivery-family composition.

(b) Delivery-family/operator map.

(c) Operator transitions.

Figure 2: Composition and structure of ChainBench-ADD delivery chains. (a) Sample counts by delivery family, split into bona fide and spoof. (b) Template-level delivery-family/operator map, where each cell shows how many templates in a family contain the operator. (c) Adjacent-operator transition matrix induced by the template inventory.

plied with Pyroomacoustics Scheibler et al. (2018), and call-path simulation jointly approximates filtering, codecs, burst loss, jitter-buffer effects and AGC. After rendering, all outputs are standardized to 16 kHz mono PCM WAV.

This hierarchy supports analysis at multiple levels, including scenario-level robustness, generalization to unseen route patterns, operator-level sensitivity, and path-wise robustness under accumulated edits. The current release contains 33 templates in total (1 direct, 6 platform-like, 10 telephony, 7 simulated replay, and 9 hybrid), which realize 26 unique ordered operator sequences and are rendered into sample-specific realized chains. Figure 2 summarizes this structure.

### 3.3 Data Construction

ChainBench-ADD is built with a staged pipeline that keeps transcript and speaker identity fixed while introducing spoofing and delivery effects in separate steps. The design follows our benchmark goal: to measure robustness to post-generation delivery artifacts rather than differences in content, speaker, or source formatting. We therefore first construct a clean parent pool of real audios, then generate spoof counterparts, and only afterward render delivery chains and export protocol-aware metadata. This ordering ensures that delivery is evaluated as a downstream factor and avoids being conflated with generator choice, transcript content, or speaker identity.

We curate bona fide speech from Common Voice 24.0 for English Ardila et al. (2020) and AISHELL-3 Shi et al. (2020) for Mandarin Chinese. Candidate utterances are screened at both the text and audio levels, removing unsuitable transcripts and low-quality recordings. Selection is speaker-centric: for each accepted speaker, a fixed number of utterances is retained after ranking candidates by properties such as usable duration, speech activity, and stable loudness. We then assign benchmark speaker IDs and create train/dev/test splits at the speaker level. The retained utterances are converted into clean masters by trimming silence, downmixing to mono, resampling to 16 kHz, normalizing loudness, and saving as 16-bit PCM WAV, yielding 18,703 bona fide clean parents.

For each bona fide clean parent, we synthesize two spoof clean parents using two of six contemporary systems: Qwen3-TTS Hu et al. (2026), CosyVoice3 Du et al. (2025), Spark-TTS Wang et al. (2025b), F5-TTS Chen et al. (2024), IndexTTS2 Zhou et al. (2025a), and VoxCPM Zhou et al. (2025b). All six systems support both English and Mandarin Chinese, spanning multiple synthesis families, including multilingual instruction TTS, zero-shot cloning TTS, expressive controllable TTS, and architecture-diverse TTS. Assignment is deterministic and load-balanced: after shuffling the bona fide parents with a fixed random seed, each parent is assigned to the two language-compatible generators with the smallest cumulative assignment counts, with deterministic tie breaking. This procedure is reproducible and keeps generator coverage approximately uniform. The speaker-disjoint split is fixed before synthesis, and all six generators are represented with similar proportions in train, dev, and test, rather than being confined to particular splits. To preserve identity while avoiding trivial prompt leakage, the target transcript is paired with a different clean utterance from the same speaker as prompt/reference audio whenever possible. Failed or invalid generations are removed. After filtering, this stage yields 37,110 spoof clean parents (about 6.1K per generator) and 55,813 clean parents in total.

Delivery chains are rendered only after the clean pool is fixed, keeping delivery orthogonal to the bona fide/spoof label. For each parent, we emit one direct control and attempt up to four sampled variants from each non-direct family by instantiating ordered waveform-domain operator chains. Within each family, template identifiers are deterministically shuffled per parent and sampled without replacement. For paired order templates, the same realized parameters are reused and only the local operator order changes, isolating ordering effects from parameter variation. Candidate renderings that fail the post-render validation contract are discarded before packaging, so the released sample counts are slightly below the nominal per-parent rendering budget. Final packaging validates audio and exports metadata that links each retained sample to its parent, family, template, variant, and operator sequence. This explicit lineage is what later enables matched local interventions and path-wise robustness analysis under shared parent context; the complete template inventory and parameter pools are given in the appendix.

### 3.4 Benchmark Tasks and Protocols

ChainBench-ADD is released as a metadata-driven benchmark for delivery-aware audio deepfake detection. From the shared metadata, we instantiate five tasks: in-chain detection, operator substitution, parameter perturbation, order swap, and delivery robustness.

The base task is in-chain detection, which performs bona fide-vs.-spoof classification separately within each delivery family. The three intervention tasks are constructed under matched parent, family, and label control. In operator substitution, paired samples have the same parent utterance, delivery family, ground-truth label, and chain length, but differ only in the operator identity at exactly one position. In parameter perturbation, paired samples differ at exactly one operator position while keeping the operator identity fixed and changing only its parameterization. In order swap, paired samples preserve the same operators and label but differ by one adjacent minimal swap in operator order. Finally, delivery robustness groups samples into lineage graphs defined by shared parent utterance, delivery family, and label, and evaluates whether correct decisions persist as additional observed delivery edits accumulate along the chain. Combined with an extra leave-one-template-out evaluation, these protocols test family-level robustness, transfer to unseen templates, sensitivity to controlled structural edits, and stability along structured delivery paths.

### 3.5 Dataset Statistics and Metadata

ChainBench-ADD contains 941,201 retained waveforms derived from 55,813 parents and 448 speakers across English and Mandarin Chinese, including 315,573 bona fide and 625,628 spoof samples. These post-validation totals are slightly below the nominal renderer budget implied by the family-level sampling policy. The speaker-disjoint train/dev/test split contains 663,363/136,027/141,811 samples from 314/67/67 speakers, respectively. The release covers five delivery families, 33 templates, and 26 ordered operator sequences, together with 182,409 operator-substitution groups, 53,234 parameter-perturbation groups, 108,245 order-swap groups, and 27,204 lineage/path groups.

The metadata are lightweight and protocol-aware. In addition to sample identifiers, transcripts, language, speaker, and provenance, they record split, generator family and model, delivery family, template and variant identifiers, the ordered operator signature and order-invariant operator multiset, realized parameters (e.g., codec, bitrate, packet loss, SNR, RT60, and source–microphone distance), and the instantiation seed. Matched-pair and lineage groups are derived deterministically from these fields under fixed grouping rules rather than manually enumerated, allowing the five protocols to be reconstructed exactly. The joint split and generator fields also recover exact per-generator counts for each split. Before export, every waveform is rechecked against target format and duration constraints, while the release manifest records split integrity, duplicate-identifier and duplicate-path checks, and parent coverage across delivery families.

### 3.6 Quality and Consistency Validation

We validate perceptual quality, matched-parent consistency, and waveform-level acoustic drift. Perceptual quality is measured with DNSMOS P.835 Reddy et al. (2022), using its P.808-based quality score and overall P.835 score (OVRL). Matched-parent consistency is evaluated on the released

Table 3: Delivery quality, matched-parent consistency, and acoustic drift by family. Dur. is the child/parent duration ratio and $|\Delta\mathrm{RMS}|$ is the absolute RMS change in dB; other abbreviations follow the text.

| Family | P808↑ | OVRL↑ | ΔWER↓ | ΔCER↓ | Spk.↑ | Dur. | \|ΔRMS\| |
|---|---|---|---|---|---|---|---|
| Direct | 3.69 | 2.91 | 0.000 | 0.000 | 1.000 | 1.000 | 0.000 |
| Platform | 3.43 | 2.76 | 0.012 | 0.011 | 0.986 | 1.005 | 0.245 |
| Telephony | 3.08 | 2.52 | 0.059 | 0.046 | 0.943 | 1.001 | 1.620 |
| SimReplay | 3.09 | 2.10 | 0.094 | 0.063 | 0.909 | 1.120 | 4.247 |
| Hybrid | 3.15 | 2.23 | 0.095 | 0.065 | 0.910 | 1.114 | 3.842 |

141,811-sample test split derived from 8,407 parents. A fixed ASR backend (Whisper-large-v3-turbo) measures child-minus-parent drift in WER and CER against the canonical transcript, while a fixed speaker backend (WavLM-base-sv) measures parent–child embedding cosine similarity. We additionally report child-to-parent duration ratio and absolute RMS change. Table 3 summarizes all family-wise means.

All families obtain P808 scores above 3.0, while OVRL is lower for telephony, simulated replay, and hybrid than for direct and platform-like delivery. The chains therefore retain usable perceptual speech quality while introducing measurable delivery-dependent variation. The same progression appears in the preservation and acoustic-drift measures: platform-like delivery remains closest to the clean parent; telephony introduces moderate ASR, speaker, and RMS changes; and simulated replay and hybrid produce the largest drift. Platform-like and telephony duration ratios remain within 0.5% of the parent, whereas simulated replay and hybrid are about 12% and 11% longer because their room-response rendering retains reverberant tails. Across the full test split, mean ΔWER/ΔCER are 0.061/0.043 and mean speaker cosine is 0.941.

For reproducibility, ASR uses Whisper-large-v3-turbo with 15-s chunks and batch size 16; speaker consistency uses a WavLM-base-sv x-vector backend with $\ell_2$-normalized embeddings and batch size 16. English hypotheses are lowercased with punctuation removed, while Mandarin WER falls back to character tokenization when whitespace segmentation is unavailable; CER is always computed over normalized characters. Together, the tabulated metrics show controlled, family-dependent acoustic variation without materially breaking the transcript and speaker correspondence required for matched-parent analysis. Row-level validation outputs and full implementation details remain in the appendix and release.

# 4 Experiments

## 4.1 Baselines and Training Setup

We evaluate AASIST Jung et al. (2021), AASIST-L Jung et al. (2021), XLS-R+SLS Zhang et al. (2024), SafeEar Li et al. (2024), and Nes2Net Liu et al. (2025), spanning graph-based, lightweight, self-supervised, content privacy-preserving, and recent efficient detector designs. We use their official implementations and modify only data loading and score export. Train/evaluation batch sizes are 24/24 for AASIST and AASIST-L, 14/8 for XLS-R+SLS (following its original configuration), 24/24 for SafeEar, and 24/16 for Nes2Net. The AASIST family is trained for 20 epochs at $10^{-3}$ learning rate; the others use 15 epochs at $10^{-6}$. All models use $10^{-4}$ weight decay. Experiments were run on an x86_64 workstation running Ubuntu 24.04.1 LTS with CUDA 12.4, eight NVIDIA RTX 3090 GPU (24 GB), two Intel Xeon Gold 6226R CPUs (32 physical cores and 64 hardware threads in total, 2.90 GHz), and 503 GiB of system memory. Baseline-specific software stacks follow their respective official implementations.

## 4.2 Evaluation Metrics and Settings

We follow the five benchmark protocols defined in Sec. 3 and introduce the evaluation setup and metrics here.

Table 4: Family-wise in-chain detection EER (%) on ChainBench-ADD. Lower is better.

| Model | Direct | Platform | Telephony | SimReplay | Hybrid | Macro |
|---|---|---|---|---|---|---|
| AASIST | 13.00 | 24.06 | 27.29 | 39.23 | 42.26 | 29.17 |
| AASIST-L | 16.89 | 24.49 | 29.30 | 41.57 | 38.79 | 30.21 |
| XLS-R+SLS | **12.92** | 11.04 | 16.01 | 20.90 | **21.11** | 16.40 |
| SafeEar | 18.14 | 27.43 | 35.65 | 37.93 | 35.94 | 31.02 |
| Nes2Net | 13.03 | **8.57** | **15.82** | **19.21** | 21.44 | **15.61** |

Table 5: Family-wise leave-one-template-out in-chain detection EER (%). Lower is better.

| Model | Platform | Telephony | SimReplay | Hybrid | Macro |
|---|---|---|---|---|---|
| AASIST | 23.52 | 28.52 | 41.66 | 41.55 | 33.81 |
| AASIST-L | 26.53 | 29.04 | 40.68 | 42.27 | 34.63 |
| XLS-R+SLS | 13.79 | 16.14 | 21.41 | **24.03** | 18.84 |
| SafeEar | 28.96 | 33.92 | 39.49 | 38.21 | 35.15 |
| Nes2Net | **11.33** | **16.08** | **21.32** | 24.09 | **18.20** |

**In-chain detection** We report family-wise EER (%) and its macro average. Leave-one-template-out evaluation excludes one template from training for each non-direct family and averages EER over the corresponding held-out templates within that family.

**Intervention robustness** For operator substitution, parameter perturbation, and order swap, we evaluate matched test pairs using the same train/dev pools as in the corresponding benchmark task. We report two primary metrics. The first is pair consistency, i.e., the proportion of pairs for which the model gives the same binary decision to both samples. The second is both-correct rate, i.e., the proportion of pairs for which both samples are classified correctly. We also report mean normalized score drift (MNSD): the mean absolute within-pair score difference divided by the score IQR of the pooled evaluation split, with denominator 1 when the IQR is at most $10^{-12}$. A single EER-defined threshold is computed on that pooled split and fixed across all pairs and lineage graphs, without family- or subgroup-specific retuning. These metrics are standardized common-operating-point diagnostics, not deployment-calibrated results. Higher is better for consistency and both-correct, and lower for MNSD.

**Path-wise delivery robustness** We evaluate delivery robustness on lineage graphs. For each lineage, the shortest observed operator signature is the reference and graph distance is the shortest path over observed atomic edits. We report the censored first-failure distance (C-FFD), which measures how far a detector remains correct before the first error; lineages with no observed failure receive $D_{\max} + 1$, where $D_{\max}$ is the largest observed graph distance. We also compute the fraction of lineages that remain correct within $k$ accumulated delivery edits, denoted by $R_k$, and define AURC-chain as the mean of $R_k$ over $k = 0, \ldots, D_{\max}$. Higher values indicate more persistent correctness; full equations are provided in the appendix.

### 4.3 Results and Analysis

**Delivery family strongly affects detection difficulty.** Table 4 reports in-chain detection EER by delivery family. The difficulty of audio deepfake detection varies systematically with the delivery family. Averaged across baselines, direct is the easiest condition (14.80 EER), followed by platform-like (19.12) and telephony (24.81), whereas simulated replay (31.77) and hybrid (31.91) are consistently more difficult. Replay-like recapture and mixed delivery effects introduce more degradation than mild redistribution. Nes2Net and XLS-R+SLS obtain the lowest macro EERs, but all baselines follow the same broad pattern across families. This reveals a stable delivery-conditioned difficulty structure obscured by pooled evaluation.

**The delivery-family ordering persists under unseen templates.** Table 5 reports leave-one-template-out in-chain detection. The overall pattern remains close to the standard in-chain results: platform-like is still the easiest family, telephony remains intermediate, and simulated replay/hybrid remain the most difficult. The family ordering therefore appears to reflect effects beyond template memorization and transfers to unseen delivery realizations.

Table 6: Matched-pair robustness under local delivery interventions. Higher is better for Cons./Both; lower is better for MNSD.

| | Operator substitution | | | Parameter perturbation | | | Order swap | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Cons.↑** | **Both↑** | **MNSD↓** | **Cons.↑** | **Both↑** | **MNSD↓** | **Cons.↑** | **Both↑** | **MNSD↓** |
| AASIST | 74.80 | 54.79 | 52.67 | 78.16 | 62.41 | 43.49 | 80.98 | 58.74 | 41.20 |
| AASIST-L | 73.75 | 52.97 | 53.10 | 77.22 | 60.88 | 34.66 | 79.19 | 55.56 | 35.16 |
| XLS-R+SLS | **87.74** | 81.07 | 20.47 | 90.36 | 85.99 | 16.48 | **90.48** | **82.62** | 16.47 |
| SafeEar | 65.25 | 47.60 | 52.90 | 67.40 | 51.29 | 45.69 | 72.56 | 51.23 | 39.85 |
| Nes2Net | 87.57 | **81.86** | **15.47** | **90.45** | **86.92** | **11.70** | 89.32 | 81.83 | **13.72** |

Table 7: Lineage robustness under accumulated delivery edits. Higher is better for all metrics.

| **Model** | **C-FFD↑** | **AURC-chain↑** | $R_0$**↑** | $R_1$**↑** | $R_2$**↑** | $R_3$**↑** |
|---|---|---|---|---|---|---|
| AASIST | 2.79 | 69.73 | 73.82 | 68.60 | 68.25 | 68.24 |
| AASIST-L | 2.61 | 65.20 | 69.21 | 64.09 | 63.76 | 63.74 |
| XLS-R+SLS | 3.47 | 86.76 | 88.39 | 86.31 | 86.18 | 86.18 |
| SafeEar | 2.46 | 61.45 | 66.97 | 59.93 | 59.45 | 59.44 |
| Nes2Net | **3.51** | **87.86** | **89.52** | **87.40** | **87.27** | **87.26** |

**Matched local interventions reveal different failure modes.** Table 6 shifts the focus from family-level variation to controlled local edits. The included three intervention tasks stress detectors in different ways. Operator substitution is the most disruptive intervention overall: averaged across baselines, it yields the lowest both-correct rate and the largest MNSD, indicating that changing operator identity perturbs detector behavior more than changing a parameter within the same operator. Parameter perturbation is the mildest intervention on average, suggesting that many detectors retain partial tolerance to within-operator variation even when they are less robust to larger structural changes in the chain. Order swap exposes a different weakness. It has the highest average pair consistency of the three tasks, yet its average both-correct rate remains well below its consistency level, especially for weaker baselines. The gap between these two quantities is largest for AASIST, AASIST-L, and SafeEar, which suggests that local reordering often preserves a model's decision without preserving its correctness. In other words, these models can be wrong on both elements of a matched pair. By contrast, XLS-R+SLS and Nes2Net show much smaller consistency–correctness gaps across all three interventions, suggesting that their invariance is more often aligned with correct discrimination rather than merely repeated error. Taken together, the matched-paired tasks help distinguish between sensitivity to operator identity, tolerance to moderate parameter variation, and vulnerability to local sequence changes.

**Robustness along delivery paths adds information beyond isolated edits.** Table 7 evaluates whether correctness persists as delivery edits accumulate along structured lineage paths. Unlike the matched-pair tasks, which isolate a single local change, this protocol tests stability under multi-step variation. The lineage metrics further separate the two strongest models. Although XLS-R+SLS and Nes2Net are similarly strong under isolated interventions, Nes2Net is consistently better in C-FFD, AURC-chain, and all reported $R_k$ values, suggesting more persistent correctness as delivery edits accumulate. The radius-wise results also show that most newly induced errors appear early: for every model, the largest drop is from $R_0$ to $R_1$, after which the curves flatten substantially, especially for the stronger baselines. Path-wise evaluation therefore shows both when failures first emerge and how quickly they propagate, complementing the view provided by isolated matched-pair tests.

## 5 Conclusion

Current audio deepfake benchmarks mainly emphasize variation on the generation side, while post-generation delivery remains underrepresented as a systematic evaluation dimension. We introduce ChainBench-ADD, a delivery-aware benchmark that provides a controlled, reproducible approximation of common delivery factors through structured chains under matched-parent control. Experiments with representative baselines show that delivery remains a major challenge: replay-like and hybrid routes are consistently harder than direct or lightly processed conditions, while protocol-aware evaluation reveals patterns missed by pooled testing. We hope ChainBench-ADD can serve as a standardized and reproducible benchmark for delivery-aware evaluation, and support future research on robust audio deepfake detection under post-generation delivery conditions.

## AI USE STATEMENT

Generative AI tools assisted manuscript writing and language revision. Their use is separate from the benchmark's speech-generation systems described in Section 3. The authors take responsibility for the final content, including claims, citations, and results.

## ETHICS STATEMENT

ChainBench-ADD supports research on robust audio deepfake detection. Speech data and voice-cloning systems carry privacy and impersonation risks; we discourage their use for deception or surveillance. Use and redistribution of the underlying data and models remain subject to the applicable licenses and access conditions. Our simulated delivery experiments do not establish reliability in high-stakes applications, and detector predictions should not be treated as conclusive evidence of authenticity.

## REPRODUCIBILITY STATEMENT

Construction and evaluation settings are described in Sections 3–4, with renderer specifications, protocol reconstruction, validation outputs, and metric definitions in Appendices A–D. We provide code, configurations, documentation, and 20 audio examples in the supplementary materials. A code repository and open dataset are available, while identifying links are omitted. The compact supplementary package excludes the full dataset and external model assets; prerequisites for full reproduction are documented in its README and construction/evaluation guides.

## REFERENCES

Hashim Ali, Nithin Sai Adupa, Surya Subramani, and Hafiz Malik. A superb-style benchmark of self-supervised speech models for audio deepfake detection. *arXiv preprint arXiv:2603.01482*, 2026.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus, 2020. URL https://arxiv.org/abs/1912.06670.

Kratika Bhagtani, Amit Kumar Singh Yadav, Paolo Bestagini, and Edward J Delp. Diffssd: A diffusion-based dataset for speech forensics. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

Ioan-Paul Ciobanu, Andrei-Iulian Hiji, Nicolae ătălin Ristea, Paul Irofti, Cristian Rusu, and Radu Tudor Ionescu. Xmad-bench: Cross-domain multilingual audio deepfake benchmark. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3109–3120, 2026.

Héctor Delgado, Giorgio Ramondetti, Emanuele Dalmasso, Gennady Karvitsky, Daniele Colibro, and Haydar Talib. On deepfake voice detection–it's all in the presentation. *arXiv preprint arXiv:2509.26471*, 2025.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*, 2021.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. SpeechFake: A large-scale multilingual speech deepfake dataset incorporating cutting-edge generation methods. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9985–9998, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.493. URL https://aclanthology.org/2025.acl-long.493/.

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 *arXiv preprint arXiv:2110.01200*, 2021.

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

Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. SafeEar: Content Privacy-Preserving Audio Deepfake Detection. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024.

Tianchi Liu, Duc-Tuan Truong, Rohan Kumar Das, Kong Aik Lee, and Haizhou Li. Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 20:12005–12018, 2025. doi: 10.1109/TIFS.2025.3626963.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023.

Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan, Jianhua Tao, Tao Wang, Shiming Wang, and Ruibo Fu. Cfad: A chinese dataset for fake audio detection. *Speech Communication*, 164: 103122, 2024.

Candy Olivia Mawalim, Yutong Wang, Aulia Adila, Shogo Okada, and Masashi Unoki. Multilingual audio deepfakes dataset for robust and generalizable detection, 2025.

Nicolas Müller, Piotr Kawa, Wei-Herng Choong, Adriana Stan, Aditya Tirumala Bukkapatnam, Karla Pizzi, Alexander Wagner, and Philip Sperl. Replay attacks against audio deepfake detection. *arXiv preprint arXiv:2505.14862*, 2025.

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *Interspeech*, 2022.

Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. Mlaad: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2024.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. *arXiv preprint arXiv:2110.01763*, 2022.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355. IEEE, 2018.

Haohan Shi, Xiyu Shi, Safak Dogan, Saif Alzubi, Tianjin Huang, and Yunxiao Zhang. Benchmarking audio deepfake detection robustness in real-world communication scenarios. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pp. 566–570. IEEE, 2025.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*, 2020.

Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. Ai-synthesized voice detection using neural vocoder artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 904–912, 2023.

Qizhou Wang, Hanxun Huang, Guansong Pang, Sarah Erfani, and Christopher Leckie. Audeter: A large-scale dataset for deepfake audio detection in open worlds. *arXiv preprint arXiv:2509.04345*, 2025a.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, 95:101825, 2026.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens, 2025b. URL https://arxiv.org/abs/2503.01710.

Yuankun Xie, Yi Lu, Ruibo Fu, Zhengqi Wen, Zhiyong Wang, Jianhua Tao, Xin Qi, Xiaopeng Wang, Yukun Liu, Haonan Cheng, et al. The codecfake dataset and countermeasures for the universally detection of deepfake audio. *IEEE Transactions on Audio, Speech and Language Processing*, 33: 386–400, 2025.

Jiangyan Yi, Chu Yuan Zhang, Jianhua Tao, Chenglong Wang, Xinrui Yan, Yong Ren, Hao Gu, and Junzuo Zhou. Add 2023: Towards audio deepfake detection and analysis in the wild. *arXiv preprint arXiv:2408.04967*, 2024.

Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, and Nicholas Evans. An initial investigation for detecting partially spoofed audio. *arXiv preprint arXiv:2104.02518*, 2021.

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*, MM '24, pp. 6765–6773, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681345. URL https://doi.org/10.1145/3664647.3681345.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*, 2025a.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*, 2025b.

Yi Zhu, Heitor R Guimarães, Arthur Pimentel, and Tiago Falk. Auddt: Audio unified deepfake detection benchmark toolkit. *arXiv preprint arXiv:2509.21597*, 2025.

## A Delivery-chain renderer specification

This appendix records the implementation details intentionally omitted from the scenario-level description in the main paper: the exact rendering budget, template inventory, parameter pools, and waveform-level operators. The released configuration and scripts are authoritative for reproducing the benchmark.

### A.1 SAMPLING POLICY

Each parent receives one direct control and a nominal budget of up to four children from each non-direct family before post-render validation. Table 8 lists the exact family inventory; failed validation removes a small number of candidate outputs from the final release. The deterministic per-parent template shuffle, sampling without replacement, and parameter-matched order pairs are described in the main paper.

Table 8: Nominal family-level budget before validation.

| Family | Templates | Children/parent |
|---|---|---|
| Direct | 1 | 1 |
| Platform-like | 6 | up to 4 |
| Telephony | 10 | up to 4 |
| Simulated replay | 7 | up to 4 |
| Hybrid | 9 | up to 4 |

### A.2 TEMPLATE INVENTORY AND PARAMETER POOLS

The released configuration contains 33 templates and 26 distinct ordered operator sequences; repeated multisets support controlled order comparisons. Table 9 gives the exact inventory.

Table 10 summarizes the fixed parameter pools and export contract.

### A.3 OPERATOR REALIZATION

Every operator is executed as a waveform transform and its realized parameters are stored in the manifest trace.

**Resampling.** ffmpeg uses the SoX resampler at precision 28. Supported one-way conversions are 16k→8k/24k/32k and 8k/24k/32k→16k; round trips apply two successive conversions through a temporary waveform.

**Band-limiting.** The narrowband profile applies second-order high/low-pass filters at 250/3400 Hz followed by companding (0.01 s attack, 0.15 s decay). The wideband profile uses 50/7000 Hz cutoffs.

**Codec and re-encoding.** AAC, Opus, GSM, $\mu$-law PCM, and A-law PCM are realized as true ffmpeg encode–decode round trips. The codec operator follows its template directly. Re-encoding uses the most recent codec in *same* mode and a different codec from $\{\text{AAC, Opus}\}$ in *cross* mode; absent prior context, simulated replay defaults to AAC and the relevant hybrid templates to Opus. GSM is encoded and decoded at 8 kHz.

**Packet loss.** Audio is divided into 20 ms frames. For target loss rate $p \leq 0.95$ and average burst length $b$, the two-state process uses

$$\begin{aligned} P(\text{bad}\to\text{good}) &= \min(1, 1/b), \\ P(\text{good}\to\text{bad}) &= \min\left(1, \frac{p\,P(\text{bad}\to\text{good})}{1-p}\right). \end{aligned} \tag{1}$$

Lost frames use repeat-fade, interpolation, or noise-fill concealment before concatenation.

**Room impulse responses.** Pyroomacoustics first derives absorption by inverse Sabine, samples valid source/microphone positions, generates and truncates the RIR, and convolves it with the waveform. A deterministic synthetic fallback uses a direct path, exponentially decaying random reflections, and a low-level noise floor. Outputs are peak-normalized after convolution.

**Additive noise.** White, pink, brown, hiss, hum, and synthetic babble are calibrated to the target SNR. Pink and brown use spectral shaping and integration, respectively; hum combines 50/100/150 Hz tones; babble sums six shifted and scaled streams. Babble, hum, and hiss use a non-stationary amplitude envelope.

Table 9: Template inventory as ordered operator blueprints before parameter instantiation.

| Family | Template ID | Ordered operator sequence |
|---|---|---|
| Direct | direct_clean | Identity chain with no operator. |
| Platform-like | aac_single | codec |
| | opus_single | codec |
| | aac_reencode | codec → reencode |
| | opus_reencode | codec → reencode |
| | aac_resample_reencode | codec → resample → reencode |
| | resample_opus | resample → codec |
| Telephony | nb_mulaw | bandlimit → codec |
| | nb_gsm | bandlimit → codec |
| | wb_opus | bandlimit → codec |
| | nb_mulaw_plr | bandlimit → codec → packet_loss |
| | nb_resample_mulaw_plr | resample → bandlimit → codec → packet_loss |
| | wb_resample_opus_plr | resample → bandlimit → codec → packet_loss |
| | wb_opus_resample_return | bandlimit → codec → resample |
| | session_nb_mulaw | call_path |
| | session_nb_gsm | call_path |
| | session_wb_opus | call_path |
| Simulated replay | rir_only | rir |
| | rir_noise | rir → noise |
| | noise_rir | noise → rir |
| | rir_reencode | rir → reencode |
| | reencode_rir | reencode → rir |
| | rir_noise_resample | rir → noise → resample |
| | resample_rir_reencode | resample → rir → reencode |
| Hybrid | opus_plr_rir | codec → packet_loss → rir |
| | rir_aac | rir → codec |
| | aac_rir | codec → rir |
| | bandlimit_codec_rir | bandlimit → codec → rir |
| | rir_bandlimit_codec | rir → bandlimit → codec |
| | rir_reencode_plr | rir → reencode → packet_loss |
| | reencode_rir_plr | reencode → rir → packet_loss |
| | resample_codec_rir | resample → codec → rir |
| | bandlimit_resample_codec | bandlimit → resample → codec |

Table 10: Fixed parameter pools and export constraints.

| Group | Released range or set |
|---|---|
| Codec bitrates | AAC {24, 32, 48}; Opus {16, 24, 32}; telephony Opus {16, 24}; re-encode {24, 32} kbps |
| Resampling | Family-specific one-way and round-trip 8/16/24/32 kHz conversions |
| Re-encoding | AAC or Opus with same/cross selection |
| Packet loss | Loss {1, 3, 5, 10}%; burst {2, 3, 5} frames; three concealment modes |
| Call path | Jitter {0, 8, 16} ms; AGC {mild, telephony} |
| Noise | Six synthetic types; SNR {30, 20, 15, 10} dB |
| RIR | RT60 {0.2, 0.4, 0.6, 0.8} s; distance {0.5, 1.0, 2.0, 3.0} m; three rooms |
| Export | Mono 16 kHz PCM WAV; duration in [1, 30] s |

**Call paths.** The composite pipeline applies profile-specific filtering/resampling, codec round trip, burst loss, frame-wise jitter-buffer simulation, and AGC/limiting. Each frame is displaced by a uniform offset in $[-J, J]$, overlap-averaged, and trimmed or padded to the original length.

# B RELEASE ARTIFACTS AND PROTOCOL RECONSTRUCTION

The package provides global and split-specific metadata CSVs, a dataset summary, dropped-row logs, and checks for speaker disjointness, duplicate identifiers/paths, format compliance, and parent coverage. The main paper lists the protocol-critical row fields. Protocol groups are derived deterministically: operator substitution fixes parent, label, family, and local context while changing one operator identity; parameter perturbation fixes identity and order while varying one parameter axis; order swap fixes the realized multiset and differs by one adjacent transposition; and lineage groups fix parent, label, and family, choose the shortest observed signature as reference, and compute graph distances over observed atomic edits. Thus the released metadata and grouping code reconstruct the benchmark without manually enumerated pair lists.

# C PRESERVATION-VALIDATION OUTPUTS

The main paper reports the validation backends, normalization rules, and family-wise DNSMOS, ASR, speaker, duration, and RMS results. The released evaluator runs from the packaged metadata (the test split by default) and writes row-level records keyed by `sample_id` and `parent_id`. Each record includes raw and normalized parent/child ASR hypotheses, WER/CER and their child-minus-parent drift, speaker cosine, duration ratio, RMS and peak-amplitude drift, the detection label, and explicit status/error fields. It also emits family-language aggregates and status counts, enabling percentile or label-conditioned analyses without rerendering. The paper pools bona fide and spoof rows because these statistics characterize delivery preservation rather than detection class.

# D BENCHMARK-SPECIFIC ROBUSTNESS METRICS

We give exact definitions for the non-standard matched-pair and lineage metrics; standard EER, AUC, accuracy, and $F_1$ follow their usual definitions.

## D.1 NOTATION AND THRESHOLDING

Let score $s_i \in \mathbb{R}$ increase with bona fide evidence and label $y_i \in \{0, 1\}$ use 1 for bona fide. For the pooled evaluation set $\mathcal{E}$, one task-level reference threshold is fixed as

$$\tau_{\text{ref}} = \begin{cases} \tau_{\text{EER}}(S_{\mathcal{E}}, Y_{\mathcal{E}}), & \text{if both classes occur,} \\ 0.5, & \text{otherwise,} \end{cases} \qquad \hat{y}_i = \mathbb{I}[s_i \geq \tau_{\text{ref}}]. \tag{2}$$

The same threshold is reused across families, matched pairs, and lineages within a task; it is not recalibrated per subgroup. Accordingly, hard-decision metrics are common-operating-point diagnostics, whereas EER/AUC and MNSD are threshold-free.

## D.2 MATCHED-PAIR METRICS

For operator substitution, parameter perturbation, and order swap, let $\mathcal{P} = \{(i_m, j_m)\}_{m=1}^{M}$ be the valid pairs whose members share a label and differ only in the targeted local intervention.

**Pair consistency rate.**

$$\text{PCR} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}[\hat{y}_{i_m} = \hat{y}_{j_m}], \tag{3}$$

PCR measures decision stability under the controlled edit, regardless of correctness.

**Pair joint accuracy.**

$$\text{PJA} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}[\hat{y}_{i_m} = y_{i_m} \ \wedge \ \hat{y}_{j_m} = y_{j_m}], \tag{4}$$

PJA requires both members to be correct.

**Mean normalized score drift.** Let $D_{\mathcal{E}}$ be the pooled-score normalization

$$D_{\mathcal{E}} = \begin{cases} \mathrm{IQR}(S_{\mathcal{E}}), & \mathrm{IQR}(S_{\mathcal{E}}) > 10^{-12}, \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

Then

$$\mathrm{MNSD} = \frac{1}{M D_{\mathcal{E}}} \sum_{m=1}^{M} |s_{i_m} - s_{j_m}|, \tag{6}$$

which measures score sensitivity independently of the hard decision.

**Symmetric misclassification risk (auxiliary).** Define sample-level 0–1 loss as

$$\ell_i = \mathbb{I}[\hat{y}_i \neq y_i]. \tag{7}$$

The corresponding pairwise quantity is

$$\mathrm{SMR} = \frac{1}{M} \sum_{m=1}^{M} \frac{\ell_{i_m} + \ell_{j_m}}{2}. \tag{8}$$

SMR is the average pairwise misclassification burden.

### D.3 Delivery-lineage metrics

A lineage groups samples with the same parent, family, and label. Each node $v$ carries an observed operator signature $\sigma(v)$; the shortest observed signature is the reference $r_g$.

**Lineage graph and node correctness.** Nodes are observed signatures; an undirected edge joins signatures that differ by one parameter perturbation, operator substitution, adjacent swap, or insertion/deletion. Let $d_g(v)$ be the shortest-path distance from $r_g$. Node correctness is

$$c(v) = \mathbb{I}[\hat{y}(v) = y(v)]. \tag{9}$$

**Censored first-failure distance.** For lineage $g$,

$$\rho_g = \min\{d_g(v) : c(v) = 0\}, \tag{10}$$

when a failure is observed; otherwise the lineage is right-censored.

Let $D_{\max} = \max_{g \in \mathcal{G}} \max_{v \in g} d_g(v)$ be the largest observed lineage distance in the evaluation set. The censored first-failure distance is then

$$\tilde{\rho}_g = \begin{cases} \rho_g, & \exists v \in g : \ c(v) = 0, \\ D_{\max} + 1, & \text{otherwise,} \end{cases} \tag{11}$$

and the reported aggregate is

$$\text{C-FFD} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \tilde{\rho}_g. \tag{12}$$

Higher C-FFD means correctness persists farther from the reference.

**Robust-at-distance.**

$$R_k = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathbb{I}[\forall v \in g : \ d_g(v) \leq k \Rightarrow c(v) = 1], \quad k = 0, 1, \ldots, D_{\max}. \tag{13}$$

$R_k$ is the fraction of lineages entirely correct through distance $k$.

**AURC-chain.**

$$\text{AURC-chain} = \frac{1}{D_{\max} + 1} \sum_{k=0}^{D_{\max}} R_k. \tag{14}$$

AURC-chain averages the complete robustness profile; higher is better.