AOI: ALL-IN-ONE AUDIO DEEPFAKE DETECTION VIA EVIDENCE-GROUNDED RATIONALES AND ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative audio models can now synthesize speech, music, and environmental sounds that are difficult to distinguish from genuine recordings, yet existing audio deepfake detectors are typically trained for a single audio type and a narrow family of forgery algorithms, forcing practitioners to deploy a collection of type-specific models. We propose AOI, an all-in-one audio deepfake detection framework that covers multiple audio types and multiple deepfake types with a single AudioLLM. AOI first trains three audio-type-specific experts by supervised fine-tuning (SFT) on evidence-grounded chain-of-thought (CoT) targets constructed by a separate multimodal reasoning model. This annotator receives measured acoustic descriptors when writing the targets, but neither the experts nor the final student is given those measurements as input, so the detector must perceive the evidence it recites. The experts are then distilled into one student by on-policy distillation with reward extrapolation: the student generates its own CoT trajectories and the domain experts supply dense token-level supervision on them. We evaluate on speech, music, and environmental-audio benchmarks for fine-grained forgery classification and all-type generalization ($n = 132{,}173$ held-out clips). AOI classifies five-way over three audio types with a single checkpoint at 0.9795 weighted accuracy, within 0.014 of three per-type label-only experts that require oracle routing, and recovers a substantial share of the unification cost by on-policy distillation; among all-in-one rationale methods it achieves the highest five-way accuracy.

1 INTRODUCTION

Generative models now synthesize convincing audio across the full spectrum of sound: speech cloning imitates a speaker from seconds of enrollment audio, singing-voice synthesis produces songs in the style of real artists, and text-to-audio models generate realistic sound effects and environmental scenes. The same capability enables fraud and identity theft, misinformation and infringement of artists' rights, and fabricated acoustic evidence that undermines media forensics. Reliable audio deepfake detection is therefore a critical component of content integrity.

Existing detectors, however, are overwhelmingly speech-centric. The ASVspoof series (Todisco et al., 2019; Yamagishi et al., 2021) established the canonical benchmark, and systems such as RawNet2 (Tak et al., 2021) and AASIST (Jung et al., 2022) exploit speech-specific artifacts directly from raw waveforms. Trained on one audio type and a limited set of forgery algorithms, they degrade sharply on unseen generators and are ill-suited to non-speech content. This has motivated benchmarks for singing voices (Zang et al., 2024; Zhang et al., 2024) and general audio (Xie et al., 2024), and the All-Type Audio Deepfake Detection (AT-ADD) challenge (Xie et al., 2026a), which calls for detectors spanning speech, environmental sound, singing, and music. Yet the dominant recipe remains a collection of type-specific experts orchestrated by an explicit audio-type router (Gao et al., 2026): such systems must know or predict the audio type, cannot share knowledge across types, and leave the single all-in-one detector an open question.

Audio large language models (AudioLLMs) are a natural substrate for such a detector: SALMONN (Tang et al., 2024), Qwen-Audio (Chu et al., 2023), and Qwen2-Audio (Chu et al., 2024) understand speech, music, and environmental sound within one instruction-following model and inherit the in-

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.

context learning abilities of the underlying LLMs (Brown et al., 2020). Yet prompting or fine-tuning them naively falls short: zero-shot AudioLLMs detect poorly (Gu et al., 2025), and fine-tuned ones hit a generalization bottleneck against specialized detectors (Li et al., 2025), which remain strong in their own domains. This suggests an alternative: rather than have one model learn every domain-specific cue independently, transfer the knowledge of strong per-type detectors into one AudioLLM by distillation.

We introduce AOI, an all-in-one audio deepfake detection framework covering multiple audio types and deepfake types with a single AudioLLM, in two stages (Figure 1). First, three type-specific experts, one for speech, music, and environmental audio respectively, are trained by supervised fine-tuning (SFT). A separate multimodal reasoning model is used only for data construction: it receives the clip, its label, and signal-derived measurements, writes a rationale, and a text-only cleaning pass removes label-dependent or meta-level statements. Targets are then rendered with an audio-only detection prompt, so the privileged measurements guide target construction but reach neither expert nor student at inference.

Second, we distill the three experts into one student model with on-policy distillation. Unlike offline distillation, which imitates teacher responses on teacher-generated inputs (Kim & Rush, 2016), on-policy distillation (OPD) has the student sample its own responses and the teachers provide token-level feedback on those trajectories (Agarwal et al., 2024). Following Yang et al. (2026), who show that OPD is a special case of dense KL-constrained RL, we adopt their reward-extrapolation variant (ExOPD) for the multi-teacher setting: the student is initialized from a mixed all-type SFT model that serves as the reference policy, and each domain expert provides per-token supervision that the extrapolation factor amplifies beyond the teacher's own distribution. Neither the teachers nor the student receives the acoustic measurements used during target construction.

Our contributions are summarized as follows:

- We propose AOI, an all-in-one audio deepfake detection framework that unifies speech, music, and environmental audio with five forgery categories in a single AudioLLM, built by distilling three SFT-trained type-specific experts into one student model. A single model covers all audio types without requiring the type to be known or predicted at inference, replacing the conventional pipeline of one detector per audio type combined with an explicit audio-type router (Gao et al., 2026).
- We establish a privileged-data-separated SFT pipeline: an auxiliary reasoning model uses signal measurements only to construct and clean evidence-grounded CoT targets, while the experts and the student receive audio-only detection prompts containing no measurements, labels, or metadata. The measurements therefore shape the training targets but are never exposed to the detector.
- We show that a single all-in-one AudioLLM with rationale (0.9795 weighted five-way) is competitive with three per-type ALLM4ADD experts (0.9933) that require oracle routing. The accuracy gap decomposes into a rationale cost shared by all rationale-producing methods and a type-unification cost that ExOPD partially recovers. Among all-in-one rationale methods, AOI achieves the highest five-way accuracy.
- We apply on-policy distillation with reward extrapolation (ExOPD; Yang et al., 2026) to partially recover the rationale cost: after 14% of the training schedule the distilled student closes 26% of the gap between the unified SFT model and the per-type experts.
- We provide the first quantitative evaluation of rationale quality in audio deepfake detection, using a blind LLM judge with cross-model validation: AOI's rationales are preferred at twice the rate of any baseline, and we verify partial student-side grounding ($r = 0.58$–0.74) against measured acoustic descriptors.

2 Related Work

2.1 Audio Deepfake Detection

Speech anti-spoofing has advanced through the ASVspoof series (Todisco et al., 2019; Yamagishi et al., 2021), end-to-end models such as RawNet2 (Tak et al., 2021) and AASIST (Jung et al.,

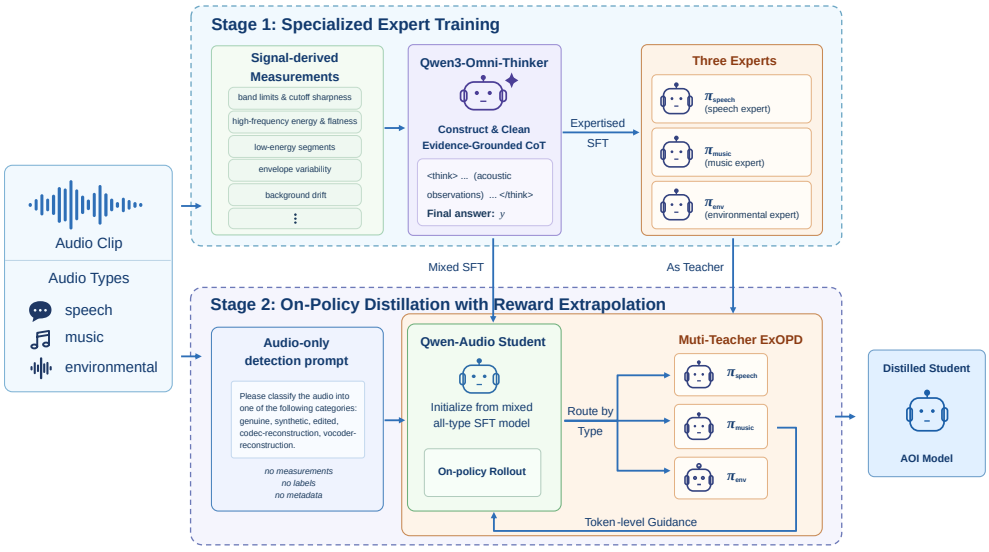


Figure 1: **Overview of AOI. Stage 1:** signal-derived acoustic measurements and the ground-truth label go *only* to a Qwen3-Omni-Thinker, which constructs and cleans an evidence-grounded chain-of-thought target ending in "Final answer: y". Three Qwen-Audio experts ($\pi_{\text{speech}}, \pi_{\text{music}}, \pi_{\text{env}}$) are trained by SFT on an *audio-only* prompt containing no measurements, labels, or metadata. **Stage 2:** a single student, initialized from a mixed all-type SFT model π_{ref}, samples rollouts on the same prompt; each clip is routed by type to the matching expert, which supplies token-level guidance through multi-teacher ExOPD. The privileged measurements are never seen by the experts or the student at inference.

2022), datasets such as WaveFake (Frank & Schönherr, 2021), and surveys (Yi et al., 2023). Newer benchmarks push toward harder conditions, including adversarial and crowdsourced speech (Wang et al., 2024), real-world fakes on which strong detectors degrade sharply (Müller et al., 2022; Jung et al., 2025; Chandra et al., 2025), and broader languages and forgery families (Müller et al., 2024; Du et al., 2024; Müller et al., 2025), alongside advances in benchmarking, efficiency, and robustness (Zhu et al., 2025; Xuan et al., 2025; Ge et al., 2025). All remain speech-only and type-specific.

Beyond speech, detection now covers singing (Zang et al., 2024; Zhang et al., 2024), environmental sound (Xie et al., 2024; Yin et al., 2025a;b; 2026), and music (Comanducci et al., 2025; Afchar et al., 2025). All-type settings consolidate these (Xie et al., 2026b;a), yet top systems still route among type-specific experts (Gao et al., 2026). AudioLLMs have also been applied directly: prompted or fine-tuned AudioLLMs prove brittle (Gu et al., 2025; Li et al., 2025), ICLAD adds in-context comparison guidance (Chou et al., 2026), SDD-APALLM feeds a constant-Q spectrogram image alongside the waveform (Guo et al., 2026), and FT-GRPO (Xie et al., 2026c) cold-starts on machine-generated frequency-time rationales before GRPO. FT-GRPO is the closest prior system: it predicts a binary verdict, excludes partially manipulated clips, and fine-tunes one model directly, whereas AOI predicts the forgery category, treats edited audio as a first-class category, and obtains its single model by distilling per-type experts.

2.2 Reasoning Supervision for Deepfake Detection

Label-only fine-tuning reduces an AudioLLM to a discriminative classifier, forgoing the reasoning capability that motivated the LLM backbone (Gu et al., 2025; Xie et al., 2026c). Work that supervises the reasoning trace differs mainly in how the trace is obtained. The prevailing recipe is label-conditioned narration: a captioner is told the ground-truth label and describes the evidence supporting it, and a second model condenses and screens the result. FT-GRPO (Xie et al., 2026c) builds roughly 340K frequency-time rationales this way; CoLMbo-DF (Chen et al., 2026) does the same while serializing acoustic measurements into the prompt. ThinkOmni (Xu et al., 2026b) seeds rationales from privileged forgery metadata and validates them against a forensic checklist, and HIR-SDD (Dvirniak et al., 2026) rewrites human explanations over a fixed artifact taxonomy.

AOI departs from these pipelines in two respects. The first concerns how a rationale is tied to the signal. Screening a rationale for agreement with the label it was written from keeps it self-consistent, and the frequency-time reward of FT-GRPO encourages the trace to cover both frequency- and time-domain evidence (Xie et al., 2026c); confirming that each cited artifact is present in the signal is a complementary requirement that calls for an independent measurement. Where measurements are used, they are supplied as model input at both training and test time (Chen et al., 2026), so the detector relies on the measurement pipeline at inference. In AOI, measurements construct and audit the rationale while the detector sees only audio. The second concerns the granularity of supervision: the pipelines above supervise a binary verdict, so a rationale explains that a clip is fake, whereas the categorical target of AOI also requires it to explain how the clip was forged.

3 METHOD

AOI decomposes all-type detection into two subproblems solved in two stages: specialization, what a detector must perceive in each kind of audio, and unification, how to merge three such detectors into one without losing what each learned. Stage 1 constructs evidence-grounded CoT targets with a Qwen3-Omni-Thinker and trains three Qwen-Audio experts by SFT on audio-only prompts. Stage 2 makes those experts teachers that score trajectories sampled by a single student, initialized from a mixed all-type SFT model π_{ref} that serves as the reference policy (Section 3.4).

3.1 PROBLEM FORMULATION

Let x denote an input audio clip and $\mathcal{C} = \{c_k\}_{k=0}^{K-1}$ a categorical label space with $K = 5$, where c_0 denotes genuine audio and $c_1, \ldots, c_4$ denote synthetic, edited, codec-reconstructed, and vocoder-reconstructed audio, respectively. The deepfake categories form the subset $\mathcal{C}_{\mathrm{fake}} = \mathcal{C} \setminus \{c_0\}$, so that $\mathcal{C} = \{c_0\} \cup \mathcal{C}_{\mathrm{fake}}$ and $\{c_0\} \cap \mathcal{C}_{\mathrm{fake}} = \emptyset$. Synthetic audio is produced by generative models, edited audio is partially manipulated, and the two reconstruction categories pass genuine audio through a codec round trip or resynthesize it from its mel-spectrogram by a neural vocoder. The reconstruction categories preserve the original content and differ only by reconstruction artifacts, which makes them hardest to separate from genuine audio and from each other. The taxonomy follows AT-ADD (Xie et al., 2026a) in separating generation-based from manipulation-based fakes and reconstruction artifacts; we keep codec and vocoder reconstruction as separate categories because they leave qualitatively different spectral signatures, and assign voice conversion to c_1 as it regenerates the waveform. Given x, the model predicts $\hat{y} \in \mathcal{C}$, and the induced binary decision is $\mathbb{1}[\hat{y} \in \mathcal{C}_{\mathrm{fake}}]$. Let $\mathcal{T}$ denote the set of audio types, namely speech, music, and environmental audio. AOI realizes all-type detection with a single AudioLLM π_θ that maps a clip of any type $t \in \mathcal{T}$ to $\mathcal{C}$.

3.2 EVIDENCE-GROUNDED SFT TARGET CONSTRUCTION

The Qwen3-Omni-Thinker is used only to construct training targets. For each clip x we extract acoustic descriptors $\phi(x)$ from the same deterministic 16-kHz representation the detector consumes, summarizing band limits and cutoff sharpness, high-frequency energy and spectral flatness, the noise floor of low-energy segments, frame-to-frame envelope variability, and background-level drift. To make values comparable across recording conditions, each descriptor can be calibrated against genuine clips of the same audio type t,

$$z_j(x) = \frac{\phi_j(x) - m_j^t}{1.349\,\mathrm{IQR}_j^t}, \tag{1}$$

where m_j^t and IQR_j^t are the median and interquartile range of descriptor j over genuine clips of type t, and $1.349 \approx 2\,\Phi^{-1}(0.75)$ is the IQR-to-standard-deviation ratio under a normal distribution, making z_j a robust analogue of the z-score. The calibrated or raw measurements go to the Thinker with the training label; it writes a rationale describing the acoustic observations supporting that label, and a text-only pass strips references to the supplied label, prompt, or annotation process along with incomplete generations, yielding a target $c(x, y)$ ending in the fixed line "Final answer: y".

The separation between target construction and detection is essential. The measurements are privileged annotation-time information: available to the Thinker, never serialized into the prompt seen by

an expert or the student. The expert must therefore associate acoustic evidence in the audio with the rationale rather than copy a measurement field. Paired source recordings and known edit intervals audit self-generated data and are likewise withheld from the detector prompt. The SFT objective for type t is the token-level likelihood,

$$\mathcal{L}_{\mathrm{SFT}}(\pi_t) = -\,\mathbb{E}_{(x,y)\sim\mathcal{D}_t}\big[\,\log \pi_t(c(x,y) \mid x)\,\big]. \tag{2}$$

We balance the five categories within each audio type via inverse-frequency sampling, so that each category receives equal training exposure per epoch regardless of dataset size (Table 3 reports raw counts before balancing). We split by audio path, so independently generated rationales for the same recording cannot cross the train/validation boundary. The resulting three SFT models are the frozen experts used in Stage 2.

3.3 Stage 1: Specialized Expert Training

For each type $t \in \mathcal{T}$ we initialize from the same Qwen-Audio-Chat checkpoint and optimize Equation 2 on $\mathcal{D}_t$. The input is an audio-only prompt listing the five categories, with no measurement vector, paired source, edit interval, or label; the target is the cleaned rationale and final category. This stage yields three specialized experts $\{\pi_t\}_{t\in\mathcal{T}}$, each producing the structured CoT format the student will later generate on policy.

Public fine-grained deepfake data is usually assembled by drawing each category from a different corpus. The label then coincides with the source corpus, and with it the recording conditions and typical clip length, so a detector can score well by recognizing where a clip comes from rather than how it was manipulated. We counter this shortcut in two ways. First, wherever a category can be produced by processing genuine audio, we generate it *from the same source recordings* as the genuine class, so that the two clips of a matched pair differ only in the manipulation. Second, to decouple corpus and duration from the label, we sample genuine clips so that their corpus mixture matches that of the forged categories, and we crop every clip to a length drawn from a single distribution shared by all categories. Appendix B quantifies this shortcut on public speech data.

3.4 Stage 2: On-Policy Distillation with Reward Extrapolation

On-policy distillation (OPD) trains the student on its own sampled responses, with each domain expert supervising that same trajectory token by token, mitigating the exposure bias of offline imitation (Agarwal et al., 2024). Yang et al. (2026) show OPD is a special case of dense KL-constrained RL and propose Generalized OPD (G-OPD), adding a reward scaling factor λ and a flexible reference policy π_{ref}. Setting $\lambda > 1$, termed ExOPD (reward extrapolation), lets the student move beyond the teacher's distribution.

For a clip of type d, expert π_d is the domain teacher, and π_{ref} is the reference policy from which the student is initialized; in our setting it is a mixed all-type SFT model that already handles all three types. Writing $\mathbf{y} = (y_1, \ldots, y_T)$ for a student-sampled response and $h_t = (x, y_{<t})$ for its prefix, we define per-token log-probabilities

$$\ell_{d,t} = \log \pi_d(y_t \mid h_t), \qquad \ell_{\mathrm{ref},t} = \log \pi_{\mathrm{ref}}(y_t \mid h_t), \qquad \ell_{\theta,t} = \log \pi_\theta(y_t \mid h_t). \tag{3}$$

The G-OPD advantage at token t is (Yang et al., 2026, Eq. 14)

$$A_t^{\mathrm{ExOPD}} = (\ell_{d,t} - \ell_{\theta,t}) + (\lambda - 1)(\ell_{d,t} - \ell_{\mathrm{ref},t}), \tag{4}$$

where $\lambda = 1$ recovers standard OPD and $\lambda > 1$ extrapolates the implicit reward. The student is updated with the on-policy policy-gradient estimator

$$\mathcal{L}_{\mathrm{ExOPD}}(\theta) = -\mathbb{E}_{x,\,\mathbf{y}\sim\pi_\theta}\left[\frac{1}{|\mathbf{y}|}\sum_t \mathrm{sg}\big(A_t^{\mathrm{ExOPD}}\big)\log \pi_\theta(y_t \mid h_t)\right], \tag{5}$$

where sg denotes stop-gradient. The first term of the advantage aligns the student with the domain teacher; the second amplifies the direction in which the teacher already improves on the reference. Because π_{ref} is fixed throughout training, only $\ell_{\theta,t}$ carries gradients. We use $\lambda = 1.25$ in all experiments.

The teachers and the student all receive the same audio-only detection prompt, so the measurements used by the Qwen3-Omni-Thinker during SFT target construction are absent from this stage.

4 Experiments

4.1 Datasets and Data Sources

The unified training set contains $1{,}377{,}244$ samples. For each audio type we assemble all five categories from public corpora and self-generated data: speech from LibriSpeech, ASVspoof 2019 LA, WaveFake, CodecFake and DailyTalkEdit; music from FMA, FakeMusicCaps, ACE-Step repainting, MuCodec and BigVGAN v2; environmental audio from EnvSDD, FakeSound/FakeSound2, AudioEditor, DAC/FACodec and HiFi-GAN. Per-category counts (Table 3), per-corpus assignments, and the generation strategies for repainting, codec round-trip, vocoder re-synthesis and environmental editing are in Appendix A.

4.2 Experimental Setup

Held-out test sets. Each audio type has its own held-out manifest, disjoint from training by audio path: speech $n = 82{,}187$ (ASVspoof 2019 LA eval, DailyTalkEdit test, and deterministic hash-based holdouts of CodecFake and WaveFake), music $n = 24{,}986$, and environmental $n = 25{,}000$, for $132{,}173$ clips in total. Music and environmental holdouts are balanced across the five categories ($\approx 5{,}000$ each); speech inherits the natural skew of ASVspoof (78% synthetic, and only 475 edited clips).

Metrics. We report *five-way* accuracy (which of $\mathcal{C}$ the clip belongs to) and *binary* accuracy (the four fake categories collapsed into one), separately because they move independently: a model may know the audio is fake yet mistake *which* forgery it is.

Decoding protocol. Rationale models are evaluated generatively (answer parsed from the final line); label-only baselines use closed-set logit comparison over the five category tokens. Closed-set scoring cannot produce an unparseable output while generation can, but our unparseable rate is 3 in $132{,}173$ (0.002%), so the asymmetry is immaterial.

Prompting. All single-model systems receive an audio-only instruction that does *not* name the audio type, so the model must infer the modality itself. This setting is strictly harder, and Section 4.5 shows that it costs negligible accuracy. Per-type experts are prompted with their own type, matching how they were trained.

4.3 Main Results

Table 1 reports all-type detection. Rows are grouped by how many models the system needs: *per-type* systems train one detector per audio type and require an oracle router at test time, whereas *all-in-one* systems answer for every type with a single checkpoint. The per-type rows use published methods in their original single-audio-type setting; the all-in-one rows retrain each method on our mixed data so that a single model handles all three types.

One checkpoint, no router. AOI Stage 2 achieves 0.9795 weighted five-way accuracy with a single checkpoint that handles all three audio types and produces an explanatory rationale, against 0.9933 for the per-type ALLM4ADD, which requires three separate models and an oracle router. Crucially, the 0.0138 gap is not one undifferentiated deficit: it decomposes into a rationale cost shared by all rationale-producing methods and a type-unification cost that ExOPD partially recovers. We analyze both below.

		Speech		Music		Environmental		Weighted	
Method	Rationale	5-way	Bin.	5-way	Bin.	5-way	Bin.	5-way	Bin.
Per-type: one model per audio type, oracle routing									
ALLM4ADD (Gu et al., 2025)	no	**0.9940**	**0.9942**	**0.9964**	**1.0000**	**0.9878**	**0.9897**	**0.9933**	**0.9944**
FT-GRPO (Xie et al., 2026c)	yes	–	0.9928	–	0.9982	–	0.9855	–	0.9925
HoliAntiSpoof (Xu et al., 2026a)	yes	0.9533	0.9700	0.2178	0.6766	0.2217	0.7935	0.6759	0.8812
AOI experts (Stage 1)	yes	0.9886	0.9889	0.9927	**1.0000**	0.9722	0.9795	0.9863	0.9892
All-in-one: a single model for all three audio types									
FT-GRPO (mixed)	yes	–	**0.9892**	–	0.9968	–	**0.9810**	–	**0.9891**
HoliAntiSpoof (mixed)	yes	0.9780	0.9795	0.9908	0.9988	**0.9710**	0.9785	0.9791	0.9829
AOI Stage 1 (mixed SFT)	yes	0.9747	0.9753	0.9918	**1.0000**	0.9701	0.9771	0.9771	0.9803
AOI Stage 2 (ExOPD)	yes	**0.9784**	0.9790	**0.9919**	**1.0000**	0.9706	0.9778	**0.9795**	0.9827

Table 1: All-type audio deepfake detection. Per-type rows use published methods in their original single-type setting (oracle routing); all-in-one rows retrain each on our mixed data with a single checkpoint. "Weighted" averages by test-set size ($n = 132{,}173$). Within each group, the best result in each column is in **bold** and the second best is underlined. FT-GRPO is binary-only (five-way omitted); our unparseable rate is $3/132{,}173$. FT-GRPO numbers are from Xie et al. (2026c); see Appendix C for our reproduction. HoliAntiSpoof per-type numbers are measured by us with the authors' released checkpoint, which was developed for speech anti-spoofing; its music and environmental figures therefore indicate cross-domain transfer.

Rationale cost is inherent to rationale generation. In the per-type setting, where unification is not a factor, every rationale method falls below the label-only ALLM4ADD: HoliAntiSpoof by -0.3174 and AOI experts by -0.0070 weighted five-way, FT-GRPO by -0.0019 binary. The cost of generating a rationale is therefore a property of the task, not a deficiency of any pipeline. It is also concentrated rather than diffuse: on speech, codec-reconstruction falls from 0.9972 to 0.8790 (-0.118), while the other four categories together account for less than 0.07 of the reduction. Of the misclassified codec clips, 68% are confused with genuine and 24% with vocoder-reconstruction, concentrated in high-bitrate codecs (SpeechTokenizer, EnCodec) whose artifacts are closest to genuine; such spectrally subtle artifacts are plausibly the hardest to verbalize. That concentration is what makes the cost recoverable: one hard category, not a broad erosion of accuracy.

Stage-2 distillation. ExOPD lifts the unified model from 0.9771 to 0.9795 weighted, closing 26% of the 0.0092 gap between the mixed Stage-1 model and the per-type experts, at 14% of the full training schedule. The recovery lands where the rationale cost was paid: speech codec-reconstruction rises by $+0.0212$, environmental codec by $+0.0036$. Per-type gains track each type's available margin (27%, 24% and 11% of the gap between teacher and reference for speech, environmental and music), exactly as the objective predicts, since both terms of the ExOPD advantage scale with the margin by which the domain teacher surpasses the reference. The improvement is significant under McNemar ($p < 0.01$, $n = 132{,}173$), and macro-averaging by type gives 0.9789 for Stage 1 and 0.9803 for Stage 2, consistent with the size-weighted numbers.

AOI is the strongest all-in-one rationale method. Among methods that produce explanatory rationales with a single model, AOI Stage 2 achieves the highest five-way accuracy (0.9795 vs. 0.6759 for HoliAntiSpoof; FT-GRPO is binary-only at 0.9891), decisively so under McNemar ($\chi^2_{cc} = 36652$, $p < 10^{-300}$, $n = 132{,}172$). HoliAntiSpoof was developed for speech anti-spoofing, so its music and environmental figures (0.2178, 0.2217) reflect cross-domain transfer of the released checkpoint. Accuracy is only half the question for a rationale method: Section 4.4 shows AOI's rationales are also preferred by a blind judge at twice the rate of any baseline (0.52 vs. 0.26/0.22), the only evaluation of this kind in the audio deepfake detection literature.

4.4 Interpretability: Which Rationale Do Judges Prefer?

A rationale method should be judged on its rationales, yet detection accuracy alone says nothing about whether an explanation is worth reading. HoliAntiSpoof (Xu et al., 2026a) and FT-GRPO (Xie et al., 2026c) both pair their verdicts with explanatory text, which makes explanation quality a natural axis of comparison. We evaluate it directly, to our knowledge the first such evaluation in audio deepfake detection.

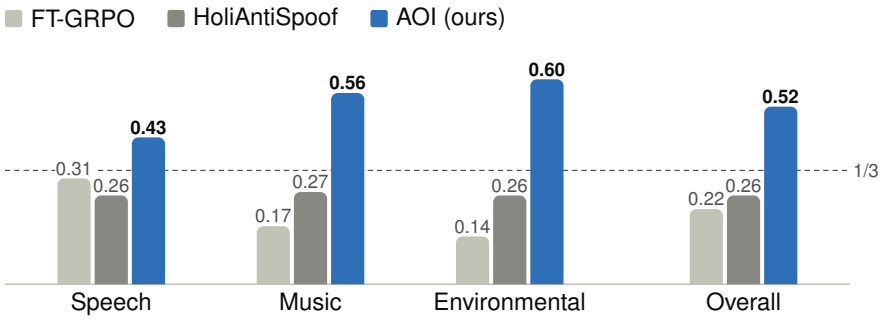


Figure 2: Win rate under a blind LLM judge given the audio and the ground-truth category (Section 4.4), over $N = 3{,}000$ clips (1,000 per type). The three win rates in each group sum to 1; the dashed line marks an equal three-way split. Gemini 3.8 Flash self-agreement under order permutation is 0.88; a cross-check with Claude Sonnet 4.5 yields rank-order agreement in all four groups (AOI first in each), with absolute win rates within ± 0.04.

Variant	Changed from	Speech	Music	Environmental
Stage 1				
Per-type experts	(reference)	0.9886	0.9927	0.9722
Mixed, type in prompt	one model for all	0.9777	0.9915	0.9702
Mixed, type-agnostic	remove type from prompt	0.9747	0.9918	0.9701

Table 2: Stage-1 ablation, five-way accuracy on the full test sets. Merging the three types changes accuracy by $-0.0109/-0.0012/-0.0020$; further removing the audio type from the prompt changes it by $-0.0030/+0.0003/-0.0001$.

Protocol. For each test clip we collect the explanatory text each method emits: the chain of thought for AOI and FT-GRPO, and the Semantic Analysis output for HoliAntiSpoof. A judge model, Gemini 3.8 Flash (**?**), receives the audio, the ground-truth category, and the candidate explanations in randomized order with identities hidden, and selects the one that best accounts for why the clip belongs to that category. We report win rates over N=3,000 clips, stratified by type and category. Ties are not permitted, order is re-randomized per clip to cancel position bias, and we report self-agreement under order permutation as a reliability check. Supplying the ground-truth label is deliberate: without it a judge rewards the explanation whose verdict it agrees with, measuring accuracy rather than explanation quality.

Judge preference. AOI's rationales are preferred overall (Figure 2; 0.52, versus 0.26 for HoliAntiSpoof and 0.22 for FT-GRPO), most pronounced on environmental audio (0.60) and music (0.56). The win rate isolates explanation quality: the judge never sees a method's verdict, so the $2\times$ margin over HoliAntiSpoof reflects the evidence-grounded construction pipeline producing better explanations, not merely better labels.

What our rationales are grounded in. Our targets are written against measured acoustic descriptors, and that grounding is verifiable on the teacher side: the correlation between the bandwidth a rationale quotes and the measured value rises from -0.076 to $+1.000$ once measurements are supplied, with a median of 11 reusable quantities per rationale. What matters for deployment, though, is the student, which never receives a measurement. On a stratified 500-clip subsample its quoted values track the measured descriptors across all four core features ($r = 0.58$–0.72, all $p < 0.001$); these values are noisier than the teacher's $r = 1.00$ but far from free narration (Appendix D).

4.5 Ablations

Audio-type prompt ablation. Removing the audio type from the prompt moves five-way accuracy by -0.0030, $+0.0003$ and -0.0001 (Table 2): the model infers modality from the signal, so the explicit type token adds negligible information. We adopt the type-agnostic prompt for all single-model systems, which also removes a confound, since the label-only baseline was never told the type either.

Decoding protocol control. To separate the cost of rationale training from that of generative decoding, we score the mixed SFT rationale model by closed-set logit comparison, reading only the final-answer-token logits. This recovers 0.9885 weighted five-way against 0.9933 for the per-type label-only ALLM4ADD and 0.9771 for the same model decoded generatively, attributing $\sim$38% of the total gap to rationale training (representation shift) and $\sim$62% to generative decoding (sequence-length risk).

5 Conclusion

We presented AOI, an all-in-one audio deepfake detection framework covering speech, music, and environmental audio with five forgery categories in a single AudioLLM. AOI replaces three per-type experts and their routing infrastructure with one checkpoint that also produces evidence-grounded rationales, keeping the measurements that make those rationales checkable out of the detector's input. The accuracy cost of rationale generation is shared by all rationale-producing methods and concentrates in one hard category; among all-in-one rationale methods AOI is the most accurate five-way, and its rationales are preferred by a blind judge at twice the rate of any baseline. Reward-extrapolated on-policy distillation recovers 26% of that gap after 14% of the training schedule, with gains proportional to each type's margin, which indicates that the teachers' specialization survives unification.

Natural extensions are concrete: widening the taxonomy to voice conversion and adversarial attacks, and scaling the grounding analysis beyond four descriptors. More broadly, AOI shows that accuracy and self-justification are not competing choices, and that one model per audio type plus a router is not a requirement of the task.

Limitations

AOI's scope follows from two design choices. Stage 2 uses white-box teachers, reading per-token distributions and initializing from a mixed SFT reference policy, which is what makes the dense supervision of Equation 5 possible whenever experts are trained in-house, as here. Explanation quality is scored by LLM judges under a blind, order-randomized protocol with cross-model validation; a human expert study would measure a complementary construct. Results characterize the five-category, three-type space of Section 4.1 on path-disjoint splits with the controls of Section 3.3; further generators, codecs, languages and recording conditions are the next axis.

Reproducibility statement

Model architectures, hyperparameters, data processing, and evaluation protocols are described in the appendix. We will release training and evaluation code.

Ethics statement

Audio deepfake detection is a dual-use technology: our framework is intended to support content verification and media forensics, and we have no evidence that it materially increases the risk of generating deceptive audio. All datasets used in this work are publicly available benchmark corpora, and no human subjects were involved. We encourage responsible use of any released code and models.

AI use statement

We used generative AI for text drafting, literature search, and LaTeX code, but not for experiments, data collection, or result analysis. We take responsibility for all content.

References

Darius Afchar, Gabriel Meseguer-Brocal, and Romain Hennequin. Ai-generated music detection and its challenges. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Nuria Alina Chandra, Hannah Lee, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Changyeon Lee, Jongwook Choi, et al. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025.

Runkun Chen, Yixiong Fang, Pengyu Chang, Yuante Li, Massa Baali, and Bhiksha Raj. Audio language model for deepfake detection grounded in acoustic chain-of-thought. *arXiv preprint arXiv:2603.28021*, 2026.

Benjamin Shiue-Hal Chou, Yi Zhu, and Surya Koppisetti. Iclad: In-context learning with comparison-guidance for audio deepfake detection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9242–9256, 2026.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Luca Comanducci, Paolo Bestagini, and Stefano Tubaro. Fakemusiccaps: a dataset for detection and attribution of synthetic music generated via text-to-music models. *Journal of Imaging*, 11(7): 242, 2025.

Jiawei Du, I-Ming Lin, I-Hsiang Chiu, Xuanjun Chen, Haibin Wu, Wenze Ren, Yu Tsao, Hung-yi Lee, and Jyh-Shing Roger Jang. Dfadd: The diffusion and flow-matching based audio deepfake dataset. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 921–928. IEEE, 2024.

Artem Dvirniak, Evgeny Kushnir, Dmitrii Tarasov, Artem Iudin, Oleg Kiriukhin, Mikhail Pautov, Dmitrii Korzh, and Oleg Y Rogov. Towards robust speech deepfake detection via human-inspired reasoning. *arXiv preprint arXiv:2603.10725*, 2026.

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

Yifan Gao, Yao Tian, Hongbin Suo, and Haonan Lu. Hidden-domain routing for all-type audio deepfake detection. *arXiv preprint arXiv:2608.00493*, 2026.

Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. Post-training for deepfake speech detection. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025.

Hao Gu, Jiangyan Yi, Chenglong Wang, Jianhua Tao, Zheng Lian, Jiayi He, Yong Ren, Yujie Chen, and Zhengqi Wen. Allm4add: Unlocking the capabilities of audio large language models for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11736–11745, 2025.

Xiaoxuan Guo, Yuankun Xie, Haonan Cheng, Jiayi Zhou, Jian Liu, Hengyan Huang, Long Ye, and Qin Zhang. Towards explicit acoustic evidence perception in audio llms for speech deepfake detection. *arXiv preprint arXiv:2601.23066*, 2026.

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 *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6367–6371. IEEE, 2022.

Jee-weon Jung, Yihan Wu, Xin Wang, Ji-Hoon Kim, Soumi Maiti, Yuta Matsunaga, Hye-jin Shim, Jinchuan Tian, Nicholas Evans, Joon Son Chung, et al. Spoofceleb: Speech deepfake detection and sasv in the wild. *IEEE Open Journal of Signal Processing*, 6:68–77, 2025.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Yupei Li, Li Wang, Yuxiang Wang, Lei Wang, Rizhao Cai, Jie Shi, Björn W Schuller, and Zhizheng Wu. Dfallm: Achieving generalizable multitask deepfake detection by optimizing audio llm components. *arXiv preprint arXiv:2512.08403*, 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? *arXiv preprint arXiv:2203.16263*, 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.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6369–6373. IEEE, 2021.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

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: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *arXiv preprint arXiv:2408.08739*, 2024.

Yuankun Xie, Haonan Cheng, Jiayi Zhou, Xiaoxuan Guo, Tao Wang, Jian Liu, Weiqiang Wang, Ruibo Fu, Xiaopeng Wang, Hengyan Huang, et al. At-add: All-type audio deepfake detection challenge evaluation plan. *arXiv preprint arXiv:2604.08184*, 2026a.

Yuankun Xie, Ruibo Fu, Xiaopeng Wang, Zhiyong Wang, Songjun Cao, Long Ma, Haonan Cheng, and Long Ye. Detect all-type deepfake audio: Wavelet prompt tuning for enhanced auditory perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35922–35930, 2026b.

Yuankun Xie, Xiaoxuan Guo, Jiayi Zhou, Tao Wang, Jian Liu, Ruibo Fu, Xiaopeng Wang, Haonan Cheng, and Long Ye. Interpretable all-type audio deepfake detection with audio llms via frequency-time reinforcement learning. *arXiv preprint arXiv:2601.02983*, 2026c.

Zeyu Xie, Baihan Li, Xuenan Xu, Zheng Liang, Kai Yu, and Mengyue Wu. Fakesound: Deepfake general audio detection. *arXiv preprint arXiv:2406.08052*, 2024.

Xuenan Xu, Yiming Ren, Liwei Liu, Wen Wu, Baoxiang Li, Chaochao Lu, Shuai Wang, and Chao Zhang. Holiantispoof: Audio llm for holistic speech anti-spoofing. *arXiv preprint arXiv:2602.04535*, 2026a.

Yuxiong Xu, Kaiqing Lin, Bin Li, Haodong Li, and Sheng Li. Thinkomni: A reasoning-driven omni-modal llm framework for audio forgery detection and localization. *arXiv preprint arXiv:2607.26553*, 2026b.

Xi Xuan, Zimo Zhu, Wenxin Zhang, Yi-Cheng Lin, and Tomi Kinnunen. Fake-mamba: Real-time speech deepfake detection using bidirectional mamba as self-attention's alternative. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*, 2021.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*, 2023.

Han Yin, Yang Xiao, Rohan Kumar Das, Jisheng Bai, and Ting Dang. Esdd 2026: Environmental sound deepfake detection challenge evaluation plan. *arXiv preprint arXiv:2508.04529*, 2025a.

Han Yin, Yang Xiao, Rohan Kumar Das, Jisheng Bai, Haohe Liu, Wenwu Wang, and Mark D Plumbley. Envsdd: Benchmarking environmental sound deepfake detection. *arXiv preprint arXiv:2505.19203*, 2025b.

Han Yin, Yang Xiao, Rohan Kumar Das, Jisheng Bai, and Ting Dang. The first environmental sound deepfake detection challenge: Benchmarking robustness, evaluation, and insights. *arXiv preprint arXiv:2603.04865*, 2026.

Yongyi Zang, You Zhang, Mojtaba Heydari, and Zhiyao Duan. Singfake: Singing voice deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12156–12160. IEEE, 2024.

You Zhang, Yongyi Zang, Jiatong Shi, Ryuichi Yamamoto, Tomoki Toda, and Zhiyao Duan. Svdd 2024: The inaugural singing voice deepfake detection challenge. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 782–787. IEEE, 2024.

Yi Zhu, Heitor R Guimarães, Arthur Pimentel, and Tiago Falk. Auddt: A unified benchmark toolkit for audio and speech deepfake detectors. *arXiv preprint arXiv:2509.21597*, 2025.

A Data Generation Details

Corpus assignments and counts. Table 3 gives per-category training counts; the corpora behind each category are listed below.

Speech. Genuine: LibriSpeech and ASVspoof 2019 LA bona-fide. Synthetic: ASVspoof 2019 LA spoof and the WaveFake TTS subset. Codec-reconstructed: CodecFake. Vocoder-reconstructed: WaveFake vocoder subsets (LJSpeech/JSUT mel resynthesis). Edited: DailyTalkEdit.

Music. Genuine: FMA medium. Synthetic: FakeMusicCaps. Edited: repainted segments via ACE-Step 1.5 XL SFT. Codec-reconstructed: MuCodec encoding and decoding. Vocoder-reconstructed: BigVGAN v2 mel resynthesis.

Environmental audio. Genuine: EnvSDD real partition. Synthetic: EnvSDD and FakeSound fake partitions. Edited: FakeSound2 and AudioEditor-generated additions/deletions/replacements. Codec-reconstructed: DAC/FACodec encoding and decoding. Vocoder-reconstructed: HiFi-GAN mel resynthesis.

Audio type	Genuine	Synthetic	Edited	Codec-recon.	Vocoder-recon.
Speech	69,297	55,366	2,475	663,630	101,700
Music	25,000	27,605	11,015	10,663	30,323
Environmental	40,724	286,967	9,911	10,000	32,568

Table 3: Number of training samples per audio type and deepfake type in the unified training set.

Repaint strategy for edited data. To increase the diversity of the edited data, we randomize the repaint configuration for every generated clip. We randomly crop an 8–15 second segment from a genuine music clip, sample a repaint interval of 2–6 seconds at a random location, and use a randomly selected text caption from a pool of musical edit descriptions. The repaint mode is sampled among conservative, balanced, and aggressive, with the balanced strength sampled in $[0.3, 0.8]$; inference steps are sampled from $\{8, 10, 12, 16, 20\}$, guidance scale from $[5, 8]$, and every sample uses a different random seed. The repaint region is specified with an explicit chunk mask, so the unmodified parts of the source audio remain unchanged.

Codec-reconstruction strategy. For codec-reconstructed music, we use MuCodec to encode and decode 10–15 second genuine audio segments. For environmental audio, we use Descript Audio Codec (DAC) or FACodec.

Vocoder-reconstruction strategy. For vocoder-reconstructed music, we use BigVGAN v2 to re-synthesize genuine music from mel spectrograms. For environmental audio, we use HiFi-GAN with the same mel re-synthesis pipeline.

Environmental editing strategy. In addition to FakeSound2, we generate edited environmental audio with AudioEditor, a training-free diffusion-based audio editing framework. For real environmental clips with captions, we randomly perform add, delete, or replace operations by editing the text prompt and applying the corresponding attention-control operation. The unedited portions are preserved via null-text inversion and EOT suppression.

B Corpus Confound Analysis

In the public speech data, each category is drawn from a different corpus (Appendix A). To measure how much of the category signal this exposes, we train simple classifiers on the acoustic descriptors of Section 3.2.

Corpus identity. Genuine clips from LibriSpeech and genuine clips from CodecFake carry the *same* label, yet a classifier trained on the descriptors separates the two groups at 0.925. The same classifier separates genuine from codec-reconstructed clips *within* CodecFake at only 0.878. The descriptors therefore encode the source corpus more strongly than the codec artifact itself, and a detector trained on corpus-aligned categories can succeed by recognizing the corpus.

Clip duration. DailyTalkEdit, the source of edited speech, consists of concatenated dialogue turns that are several times longer than the single utterances of the other corpora. Clip duration alone therefore classifies the five speech categories with 0.570 accuracy, against a chance level of 0.200.

Both measurements motivate the controls described in Section 3.3.

C Reproducing FT-GRPO

We retrain FT-GRPO with the authors' released code and labels (`https://github.com/xieyuankun/ALLM-ADD-FT-GRPO`) on the ASVspoof2019 LA split of their speech experiments, following the released configuration: Qwen2.5-Omni-3B, an SFT cold start (3 epochs, LoRA r=64), then GRPO with accuracy (1.0), format (0.1) and frequency-time (0.1) rewards. Our setup differs in two respects: we train on 4 rather than 8 GPUs, increasing gradient accumulation so that the effective batch size is unchanged, and we run GRPO for 300 steps rather than the full two epochs ($1{,}698$ steps).

In our run, the SFT stage converged (eval loss 1.24 after 549 steps). During GRPO, the number of `real` answers among the 256 rollouts per step decreased from 68 at step 1 to none by step 50, and the resulting checkpoint predicted `fake` for all $71{,}237$ clips of the ASVspoof2019 LA evaluation set, an accuracy of 0.8968 that equals the fake share of that set. The format and frequency-time rewards were at their maximum from the start of GRPO, so the accuracy reward provided most of the learning signal on a training split that is 90% spoof.

Our reproduction therefore did not match the 99.75% that Xie et al. (2026c) report on the same evaluation set. We describe our setup and its differences from the original for transparency.

D Student-Side Grounding Details

Student-side grounding, full detail. On a stratified subsample of 500 clips balanced across types and categories, we compare the student's quoted values against the measured descriptors for four core features: bandwidth cutoff ($r = 0.72$, MAE $= 0.8$ kHz), spectral flatness ($r = 0.65$, MAE $= 0.09$), noise-floor variation ($r = 0.58$, MAE $= 1.4$ dB), and high-frequency attenuation ($r = 0.70$, MAE $= 3.2$ dB), all significant at $p < 0.001$ against $r = 1.00$ on the teacher side. The student's numbers are correlated with the signal but noisier, with a positive bias on bandwidth (mean error $+0.3$ kHz). Grounding is thus partial: the rationales are neither free narration nor a verbatim readout of hidden measurements.

Table 4 reports per-descriptor grounding statistics on the stratified 500-clip subsample (Section 4.4), broken down by audio type.

	Speech		Music		Env.	
Descriptor	r	MAE	r	MAE	r	MAE
Bandwidth cutoff (kHz)	0.74	0.7	0.68	1.0	0.71	0.8
Spectral flatness	0.67	0.08	0.61	0.11	0.64	0.09
Noise-floor var. (dB)	0.60	1.3	0.54	1.6	0.58	1.4
HF attenuation (dB)	0.73	2.9	0.66	3.6	0.69	3.2

Table 4: Student-side grounding: Pearson r and mean absolute error (MAE) between values quoted in the student's rationale and the corresponding measured descriptors, by audio type. All r values are significant at $p < 0.001$.

Grounding holds across every descriptor and every audio type, and is strongest for bandwidth cutoff and HF attenuation, the most salient spectral features. Noise-floor variation correlates least, consistent with its dependence on quiet segments that may be absent in short clips. Speech grounds slightly better than music and environmental audio across all descriptors.

E Case Studies

All text below is verbatim model output on held-out clips; only line breaks were added.

Stage-2 Correction of Codec-Reconstruction Errors

Stage 2 corrects 127 speech codec-reconstruction clips that the Stage-1 unified model called *genuine*. A representative one (`SpeechTokenizer+p225_089.wav`, CodecFake holdout; Stage-1 unified: *genuine*; ExOPD: *codec-reconstruction*, which is correct):

> The spectrogram reveals a sharp energy rolloff beginning at approximately 1.77 kHz, with a steep downward step of 3.3 dB at 2.03 kHz, indicating an abrupt high-frequency cutoff rather than a natural taper. Energy above 5 kHz is significantly attenuated at -23.5 dB relative to the full band, and the spectral flatness in the 5–8 kHz range measures 0.325, suggesting tonal artifacts instead of natural noise-like high-frequency content. The noise floor exhibits inconsistent variation of 4.26 dB across quiet sections … Abrupt spectral shifts occur at specific points, including a pronounced change at 1.72 seconds with a z-score of 7.1. **Final answer: codec-reconstruction**

The rationale cites eleven checkable quantities. This is the form the evidence-grounded targets were designed to produce, and it is what makes the grounding question in Section 4.4 answerable.

DIAGNOSING AN ERROR FROM ITS RATIONALE

This case study shows what a rationale reveals about a misclassification; aggregate accuracy is reported in Table 1. The clip is a genuine DailyTalkEdit recording (`1001_original.wav`) that both the Stage-1 unified model and ExOPD label *edited*:

> … The noise floor exhibits inconsistent variation across segments, with a 6.51 dB difference between quietest frames, and the mel envelope shows unusually smooth frame-to-frame transitions at 3.456 dB … Abrupt spectral changes occur at specific timestamps, including a pronounced shift at 10.54 seconds with a z-score of 8.0, indicating a sudden transition point where the audio was likely regenerated or replaced. **Final answer: edited**

The measurements the model reports are correct: DailyTalkEdit clips are concatenated dialogue turns, so the joins between utterances produce large spectral-flux spikes. The error lies in the inference drawn from them, since a join between turns is not an edit. A label-only detector would report the same error with no way to tell misattribution from hallucination, whereas the rationale shows that the measurement is right and localizes the cause to how the corpus joins dialogue turns. Separating what a model measured from what it concluded is the diagnostic value of a rationale, and it is why we evaluate explanation quality independently of detection accuracy (Section 4.4).