GenTraceBench: A Benchmark for Tracing Audio Deepfakes Across Pre- and Post-training Stages

Anonymous ACL submission

Abstract

Modern text-to-speech (TTS) models increasingly rely on foundation pretraining followed by post-training adaptation, creating new challenges for audio deepfake detection and attribution in the wild. Prior benchmarks mainly test against fixed generators and thus underestimate the impact of adaptation-induced shifts. We present GenTrace, a benchmark that tracks TTS evolution from foundation pretraining to diverse adaptation strategies, with controlled prompts and speakers to isolate model-induced differences (16 variants, 49,728 synthesized utterances). Using GenTrace, we find that alignment-based adaptation typically preserves detection accuracy, while architecture and pretraining data have a substantially larger effect on attribution performance. GenTrace supports reproducible evaluation of detection and attribution robustness under realistic model adaptation scenarios. GenTrace will be publicly released upon acceptance.

1 Introduction

The rapid advancement of generative speech technologies has significantly lowered the barrier to producing highly realistic synthetic audio, raising growing concerns about audio deepfakes in real-world scenarios. As text-to-speech (TTS) models continue to improve in fidelity, controllability, and accessibility, reliable detection and attribution of synthetic speech have become increasingly important for forensic analysis, media authentication, and trustworthiness assessment.

1.1 Evolution of Generative Speech Paradigms

Recent TTS systems increasingly follow a two-stage workflow (Zhang et al., 2023; Chen et al., 2024; Zhang et al., 2025a): large-scale foundation pretraining to learn general acoustic representations, followed by post-training adaptation to optimize controllability and subjective quality. In this paradigm, foundation models (Wang et al., 2023; Anastassiou et al., 2024; Du et al., 2024) provide a strong initialization, while adaptation methods such as supervised fine-tuning (SFT) (Yang et al., 2023) and preference-based alignment (Rafailov et al., 2024) (e.g., DPO/GRPO) further refine prosody, style, and instruction following.

1.2 Critical Gap in Evaluation Benchmarks

However, evaluation benchmarks for audio deepfake detection have not kept pace with this shift. Most existing benchmarks (Yamagishi et al., 2021; Liu et al., 2023; Cai et al., 2024) were designed around earlier, monolithic generators and provide limited coverage of the post-training adaptations that dominate current model lifecycles. Detection systems are typically evaluated on the same type of generators seen during training, rather than on adapted models used in real attacks, so their robustness to adaptation-induced distribution shifts is often unclear (Müller et al., 2022; Frank and Schönherr, 2023). In particular, SFT can introduce new behaviors and artifacts, while RL-based alignment can induce subtle distribution changes; both may affect detection and attribution in ways that are not captured by conventional benchmarks.

1.3 GenTrace: Benchmark Design Overview

To address this gap, we introduce GenTrace, a benchmark designed to evaluate audio deepfake detection and attribution under realistic TTS model evolution. GenTrace systematically covers key stages of modern TTS development, from foundation pretraining to diverse post-training adaptations, including supervised fine-tuning, reinforcement-learning-based alignment, and multi-stage optimization. By evaluating multiple representative TTS architectures under controlled prompts and speaker identities, GenTrace enables a focused analysis of how different adaptation strategies affect detection robustness and attribution reliability.

2 GenTrace Benchmark Construction

To reflect the contemporary transition from foundation pretraining to post-training adaptation, we construct GenTrace, a benchmark designed to study how discriminative features used for audio deepfake detection and attribution change across the model lifecycle.

2.1 Model Selection and Adaptation Framework

GenTrace encompasses five representative state-of-the-art speech generation architectures (CosyVoice2 (Du et al., 2024), F5-TTS (Chen et al., 2025), FlexiVoice (Anonymous, 2025), MaskGCT (Wang et al., 2025b), and Vevo2 (Zhang et al., 2025b)), evaluated across both pretraining and post-training phases. These models span diverse acoustic modeling paradigms, including LLM-conditioned flow matching (CosyVoice2), diffusion-based flow matching (F5-TTS), hybrid autoregressive token modeling with flow-matching vocoders (FlexiVoice and Vevo2), and masked generative modeling for acoustic reconstruction (MaskGCT).

Pretraining. Most models utilize foundation versions pretrained on the extensive Emilia dataset (He et al., 2024), except CosyVoice2, which employs its official large-scale dataset. Additionally, to exploit Vevo2's singing voice synthesis capabilities, we include a variant pretrained on the SingNet (Gu et al., 2025) dataset for cross-domain analysis.

Post-training adaptation. We implement three dominant adaptation settings that reflect practical optimization paths: (i) **general alignment** via DPO on the INTP dataset (Zhang et al., 2025a) for all models to enhance speech intelligibility; (ii) **modality expansion** for CosyVoice2 via SFT on NVSpeech (Liao et al., 2025) to incorporate non-linguistic vocalizations (e.g., laughter); and (iii) **multi-stage evolution** for Vevo2 through sequential DPO→GRPO alignment, and for FlexiVoice through S1 (multi-modal DPO)→S2 (decoupling GRPO)→S3 (instruction GRPO).

2.2 Data Generation Protocol

We generate speech using the widely adopted Seed-TTS evaluation corpus (Anastassiou et al., 2024) as prompts, which provides diverse speakers and 3,108 utterances (1,088 English and 2,020 Chinese). A key design choice is strict variable control: the same speaker identities and text prompts are used across all model variants, so that observed differences can be attributed to model and adaptation factors rather than changes in content or speaker.

2.3 Dataset Partitioning

The source corpus maintains a 2:1 Chinese-to-English ratio and is split into training (1,035), validation (517), and test (1,556) sets, with identical speaker identities and text content preserved across variants for controlled comparison. GenTrace covers 16 model variants and 49,728 synthesized utterances, together with *bona fide* human speech.

3 Experimental Setup

3.1 Task Formulation

We consider three complementary tasks: **binary detection**, **closed-set attribution**, and **deepfake verification**.

Binary detection. The goal of detection is to distinguish between *bona fide* human speech and synthetic speech generated by TTS models. This task is formulated as a binary classification problem, where each audio sample is labeled as either bona fide or spoof (Wu et al., 2015).

Closed-set attribution. Closed-set attribution is formulated as a multi-class classification problem over a fixed set of known TTS generator families. Given a synthetic audio sample, the model predicts which generator family produced it (Anastassiou et al., 2024). While useful for analyzing discriminative features among known generators, this setting assumes that all possible generators are observed during training and thus does not generalize to unseen or emerging models.

Deepfake verification. Deepfake verification determines whether two audio samples originate from the same TTS generator (Wang et al., 2025a). Unlike closed-set attribution, this open-set formulation does not assume that all generators are observed during training, enabling evaluation of fingerprint stability under post-training adaptation (Klein et al., 2025).

3.2 Detection and Attribution Model

We use a pretrained Wav2Vec2-BERT (Chung et al., 2021) model as the backbone to extract high-level acoustic features from raw audio. During training, the backbone is fine-tuned together with the downstream classifiers. For detection, a linear classifier

is trained on top of the Wav2Vec2-BERT representations to distinguish bona fide human speech from synthetic speech. For attribution, a linear classifier is trained to predict the source TTS model family under a closed-set setting.

3.3 Training Protocol

Models are trained on the training split of GenTrace and selected using the validation split. All hyperparameters are fixed across experiments to ensure comparability. To evaluate robustness under realistic deployment scenarios, we adopt a train-on-foundation, test-on-adapted protocol. Detection and attribution models are trained exclusively on samples generated by foundation TTS models and bona fide speech, and are then evaluated on samples from unseen adapted variants. This setting simulates practical conditions in which detection systems are deployed against adapted generators that were not available during training.

3.4 Evaluation Metrics

For detection, we report test-set classification accuracy. For closed-set attribution, we report top-1 accuracy across generator classes, complemented by confusion matrices and feature-space visualizations. For deepfake verification, we report the Equal Error Rate (EER).

4 Experimental Results and Analysis

We evaluate robustness to post-training adaptation using quantitative results (Figure 1), with additional feature-space visualizations provided in Appendix B (Figures 3 and 4).

4.1 Impact of Adaptation Stages on Detection and Attribution Stability

RL-based alignment largely preserves detection and attribution performance, whereas data/domain changes and SFT produce clear feature shifts that degrade generalization. DPO/GRPO leads to only negligible changes in both detection and attribution (e.g., Vevo2 attribution changes from 92.16% to 91.07%, and F5-TTS detection from 99.81% to 100%). The t-SNE plot in Appendix B (Figure 3) shows base models and their aligned counterparts forming tightly overlapping clusters, suggesting that preference optimization primarily affects sampling behavior without substantially restructuring the latent acoustic representation.

In contrast, changing Vevo2's pretraining corpus from Emilia to SingNet reduces attribution accuracy by 8.61%, and the corresponding cluster exhibits a distinct movement in feature space. For CosyVoice2, SFT on NVSpeech reduces detection accuracy to 87.28%, consistent with modality expansion altering internal decoding patterns more substantially than alignment.

4.2 The Quality–Detectability Trade-off

Speech fidelity and detectability exhibit a non-monotonic, architecture-dependent relationship; improving perceptual quality can either reduce or preserve detectability depending on the generator. CosyVoice2 achieves high fidelity (WER 1.64%, SIM-o 0.720) but shows the lowest base-model detection accuracy (89.01%). After DPO, WER improves to 1.43% while detection decreases to 87.66%, implying that optimization may reduce detectable artifacts for this architecture. For F5-TTS, DPO reduces WER from 4.63% to 1.66% yet detection saturates at 100%, suggesting that perceptual improvements do not necessarily translate to improved evasion and may instead consolidate regularities exploitable by detectors.

4.3 Deepfake Attribution Analysis

Closed-set attribution is primarily determined by architecture-level signatures, but attribution can become ambiguous when models share priors (data/vocoder) and thus overlap in feature space. The t-SNE visualization in Appendix B (Figure 3) shows five model families forming distinct macro-clusters despite adaptation. A notable exception is the strong confusion between FlexiVoice and Vevo2: 99% of FlexiVoice errors are misclassified as Vevo2. This pattern aligns with their entangled clusters and high embedding cosine similarity in the embedding similarity heatmap (Appendix Figure 4), consistent with shared lineage (same Emilia pretraining data and Flow-Matching/Vocos backend).

4.4 Quantifying Fingerprint Drift via Multi-shot Verification

While detection and attribution tasks evaluate general decision boundaries, they do not fully quantify the **structural integrity** of model-specific fingerprints under adaptation. To probe the nature of adaptation-induced drift, we conducted a **Multi-shot Verification** experiment (Fig. 2), measuring the Equal Error Rate (EER) between base models and their adapted variants using prototype vectors averaged from N samples.

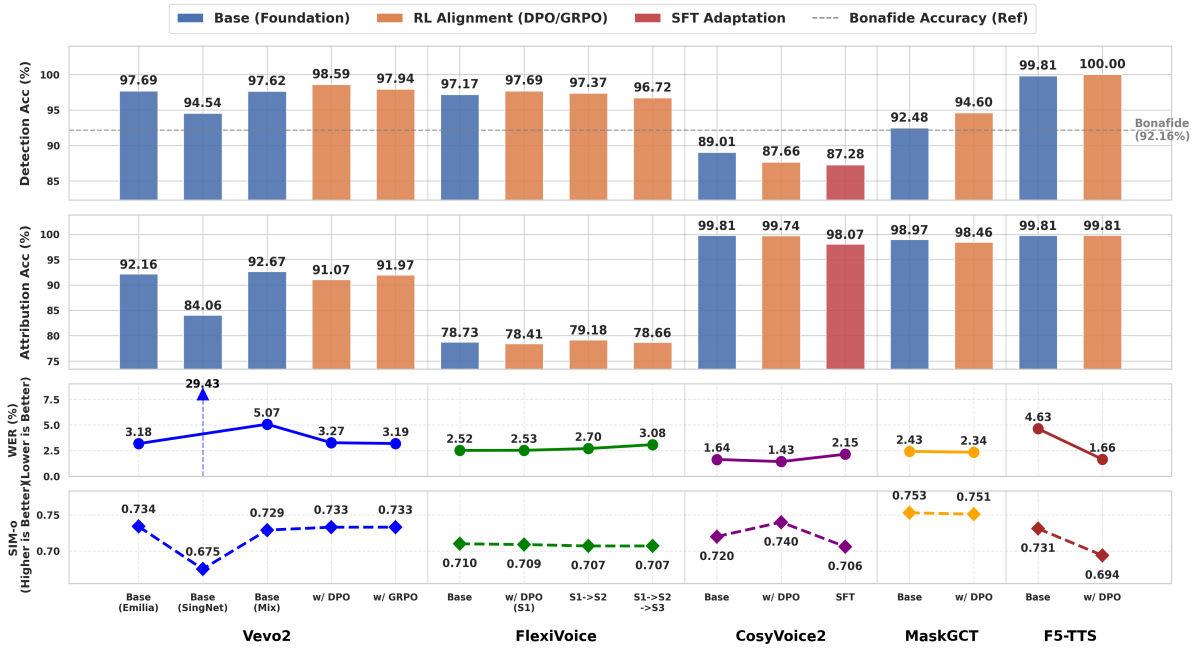

Figure 1: Overall comparison of detection accuracy, attribution accuracy, and speech quality (WER and SIM-o) across foundation models and adapted variants. Blue bars denote foundation models, orange bars denote RL-based alignment (DPO/GRPO), and red bars denote SFT adaptation. The dashed line indicates the reference accuracy on bona fide speech.

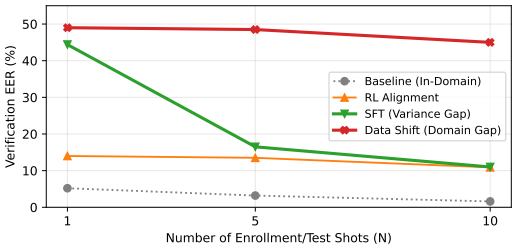

Figure 2: **Fingerprint Identity Verification using Multi-shot Prototypes.** We evaluate the Equal Error Rate (EER) under varying numbers of enrollment/test shots ($N = \{1, 5, 10\}$). While *SFT (Green line)* shows a sharp EER drop as N increases (indicating high variance but recoverable identity), *Data Shift (Red line)* remains persistently high, signifying an irreversible domain gap.

Variance-Driven Drift (The SFT Case): For CosyVoice2, SFT adaptation initially results in a high EER (44.4%) at $N = 1$. However, increasing the prototype shots to $N = 10$ dramatically reduces the EER to **11.0%**. This sharp decline suggests that SFT primarily introduces increased **intra-class variance** into the feature space. By averaging multiple samples, the core architectural identity can be effectively recovered.

Structural Domain Gap (The Data Shift Case): The ineffectiveness of averaging for Vevo2-SingNet suggests that the domain shift (Speech-to-Singing) maintains angular collinearity with the base model in the cosine space. This confirms that the verification model primarily locks onto the vocoder-induced artifacts (which are identical across domains) rather than prosodic variations. The high EER is therefore a result of the model's robustness to domain changes, paradoxically leading to "verification failure" when differentiation is desired.

5 Conclusion

GenTrace establishes a comprehensive benchmark addressing the critical gap in evaluating deepfake detection systems against modern adapted speech generation models. Our findings reveal that while alignment techniques minimally affect architectural fingerprints, pretraining data composition and fundamental model architecture constitute dominant determinants of attribution performance. The dataset and associated analyses provide essential insights for developing robust detection frameworks capable of addressing the evolving landscape of synthetic speech generation. Future work will expand coverage to additional adaptation techniques and investigate the transferability of detection features across diverse generative speech paradigms.

6 Limitations

Our study has three main limitations. First, due to computational constraints, we explored downstream adaptations (SFT and Data Shift) on representative models (CosyVoice2 and Vevo2) rather than all architectures. Future work should expand this to a full factorial analysis to verify the universality of "adaptation drift". Second, GenTrace currently focuses on clean, high-fidelity audio to isolate model-specific artifacts. The impact of transmission channels (e.g., MP3 compression, telephony) remains to be investigated. Finally, while we cover two major languages (English/Chinese), expanding to multilingual scenarios remains a necessary step for global forensic applicability.

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, and 27 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *Preprint*, arXiv:2406.02430.

Anonymous. 2025. TTS can speak in any style with any voice. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

Zhixi Cai and 1 others. 2024. Voice cloning and deepfake detection: A survey. *arXiv preprint arXiv:2401.12345*.

Tong Chen, Zhaoxi Zhang, Xijun Liu, and 1 others. 2024. Speechalign: Aligning speech generation to human preferences. In *ICASSP*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *Preprint*, arXiv:2108.06209.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Joel Frank and Lea Schönherr. 2023. Wavefake: A data set to facilitate audio deepfake detection. In *NeurIPS Datasets and Benchmarks Track*.

Yicheng Gu, Chaoren Wang, Junan Zhang, Xueyao Zhang, Zihao Fang, Haorui He, and Zhizheng Wu. 2025. Singnet: Towards a large-scale, diverse, and in-the-wild singing voice dataset. *Preprint*, arXiv:2505.09325.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. *Preprint*, arXiv:2407.05361.

Nicholas Klein, Hemlata Tak, and Elie Khoury. 2025. Open-set source tracing of audio deepfake systems. *Preprint*, arXiv:2507.06470.

Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang, and Zhizheng Wu. 2025. Nvspeech: An integrated and scalable pipeline for human-like speech modeling with paralinguistic vocalizations. *Preprint*, arXiv:2508.04195.

Bo Liu, Junichi Yamagishi, and 1 others. 2023. Asvspoof 5: the forthcoming challenge on detecting spoof of diverse sources. In *Interspeech*.

Nicolas M Müller, Franziska Dieckmann, and 1 others. 2022. Does audio deepfake detection generalize? In *Interspeech*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Chengyi Wang, Sanyuan Chen, Yu Wu, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Li Wang, Junyi Ao, Linyong Gan, Yuancheng Wang, Xueyao Zhang, and Zhizheng Wu. 2025a. Audio deepfake verification. *Preprint*, arXiv:2509.08476.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Shunsi Zhang, Xueyao Zhang, and Zhizheng Wu. 2025b. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. In *13th International Conference on Learning Representations, ICLR 2025*.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilci, Md. Sahidullah, and Aleksandr Sizov. 2015. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *INTERSPEECH 2015 16th*

Annual Conference of the International Speech Communication Association, pages 2037–2041, France. International Speech Communication Association.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, and 1 others. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop*.

Dongchao Yang, Songxiang Liu, Rongjie Huang, and 1 others. 2023. Instructtts: Modelling intentions for speech synthesis via instructors. In *Interspeech*.

Dong Zhang, Shimin Yu, Xijun Liu, and 1 others. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *EMNLP*.

Xueyao Zhang, Yuancheng Wang, Chaoren Wang, Ziniu Li, Zhuo Chen, and Zhizheng Wu. 2025a. Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12251–12270, Vienna, Austria. Association for Computational Linguistics.

Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu. 2025b. Vevo2: A unified and controllable framework for speech and singing voice generation. *Preprint*, arXiv:2508.16332.

A Detailed Experimental Results

Here, we present the detailed experimental results for deepfake detection and attribution across various model variants and adaptation methods. Table 1 summarizes the performance metrics including detection accuracy, attribution accuracy, WER, and SIM-o for all evaluated models.

Table 1: Performance of Deepfake Detection and Attribution (Accuracy; WER; SIM-o).

Model	Detection Acc(%)↑	Attribution Acc(%)↑	WER(%)↓	SIM-o↑
Bona fide	92.16	NA	NA	NA
Part I: Post-training Adaptation				
Vevo2 (Pre-training)	97.69	92.16	3.18	0.734
w/ DPO	98.59	91.07	3.27	0.733
w/ GRPO	97.94	91.97	3.19	0.733
FlexiVoice (Pre-training)	97.17	78.73	2.52	0.710
w/ DPO (S1)	97.69	78.41	2.53	0.709
S1→S2	97.37	79.18	2.70	0.707
S1→S2→S3	96.72	78.66	3.08	0.707
CosyVoice2	89.01	99.81	1.64	0.720
w/ DPO	87.66	99.74	1.43	0.740
SFT	87.28	98.07	2.15	0.706
MaskGCT	92.48	98.97	2.43	0.753
w/ DPO	94.60	98.46	2.34	0.751
F5-TTS	99.81	99.81	4.63	0.731
w/ DPO	100.00	99.81	1.66	0.694
Part II: Pre-training Dataset (Vevo2)				
Vevo2 (Emilia+SingNet)	97.62	92.67	5.07	0.729
w/o SingNet	97.69	92.16	3.18	0.734
w/o Emilia	94.54	84.06	29.43	0.675

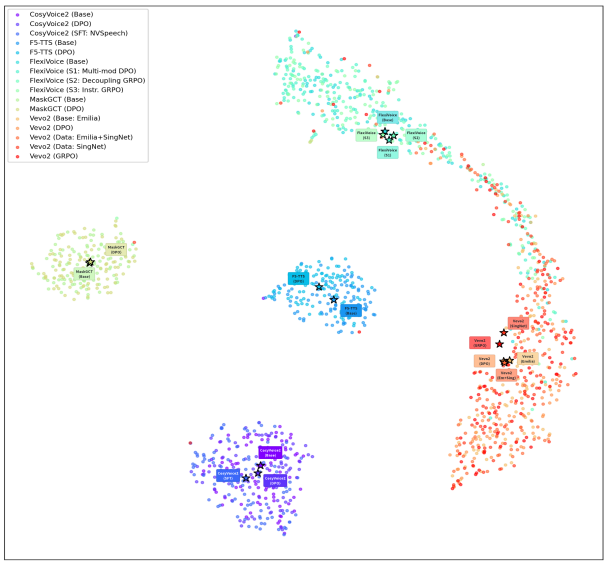

Figure 3: t-SNE clustering visualization of model features.

B Additional Visualization Analyses

To complement the quantitative evaluations, we further analyze the learned acoustic representations through feature-space visualizations, including t-SNE projections and an embedding similarity heatmap. These visual analyses provide intuitive insights into the geometric structure of model-specific acoustic fingerprints, as well as the relationships among different architectures and adaptation strategies

B.1 Acoustic Fingerprint Landscape

Figure 3 presents a two-dimensional t-SNE projection of high-dimensional Wav2Vec2-BERT embeddings extracted from synthetic speech. Several salient structural patterns emerge:

Architecture-Level Clustering. Models built upon distinct foundational architectures—such as CosyVoice2 (purple), F5-TTS (blue), and MaskGCT (light green)—form clearly separated clusters in the embedding space. This pronounced spatial segregation indicates that the underlying generative architecture constitutes the dominant factor shaping the global acoustic fingerprint, outweighing the effects of downstream adaptation.

Stability under RL-based Alignment. Within each architectural cluster, the base model (solid star markers) and its RL-aligned variants (hollow stars) are tightly co-located. This observation provides visual confirmation that preference-based alignment methods (e.g., DPO/GRPO) induce only minor perturbations in the latent acoustic distribution, pre-

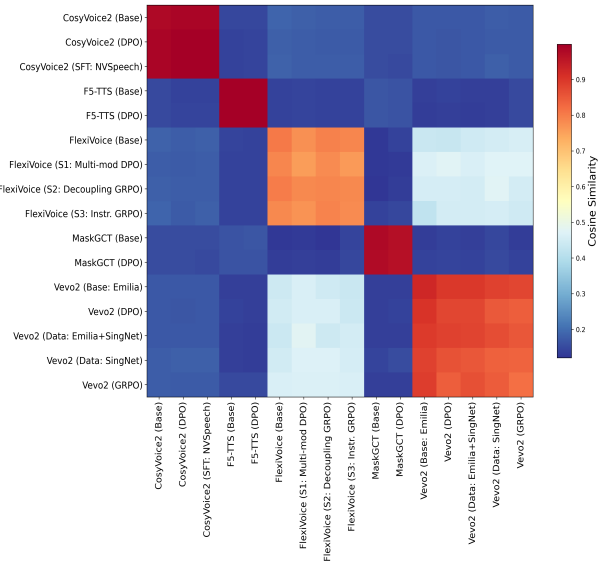

Figure 4: Embedding similarity analysis.

serving the core fingerprint of the generator.

Trajectory of Data-Induced Shift. In contrast, the Vevo2-SingNet variant exhibits a pronounced displacement from the main Vevo2 cluster, forming a long, continuous trajectory extending to the right of the embedding space. This drift reflects the substantial domain shift from speech-centric pretraining (Emilia) to singing-oriented data (SingNet), visually corroborating the elevated attribution error observed in the quantitative experiments.

Visualizing Lineage Entanglement. Notably, the FlexiVoice (teal) and Vevo2-Emilia (orange) clusters appear adjacent with partially overlapping boundaries. This proximity provides a geometric interpretation of the lineage entanglement phenomenon: models that share pretraining data and vocoder backends develop highly similar macro-level acoustic representations, leading to ambiguity in attribution.

B.2 Embedding Similarity Heatmap

Figure 4 depicts the pairwise cosine similarity matrix among model embeddings, where darker red indicates higher similarity.

Intra-Model Consistency. The prominent diagonal blocks confirm strong intra-model similarity: samples generated by the same architecture—regardless of alignment or minor adaptations—exhibit highly consistent representations, reinforcing the notion of stable architectural fingerprints.

Quantifying Attribution Ambiguity. The most striking off-diagonal pattern corresponds to the high-similarity block between FlexiVoice and Vevo2. The deep red coloration in this cross-region indicates near-collinearity of their feature vectors, quantitatively substantiating the confusion observed in closed-set attribution. This result supports our conclusion that shared macro-priors, particularly pretraining data and vocoder design, can effectively collapse multiple generators into a unified super-class fingerprint, thereby necessitating finer-grained or architecture-aware discriminators for reliable attribution.