# ENVTTS: TOWARDS INDEPENDENT SPEAKER AND ENVIRONMENT CLONING IN TEXT-TO-SPEECH

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

## ABSTRACT

Zero-shot text-to-speech (TTS) has made substantial progress in cloning unseen speakers from reference speech. However, the reference speech also captures acoustic characteristics of its recording environment, making speaker identity and environment difficult to control independently. **Environment-controllable TTS (EnvTTS)** uses separate speaker and environment references to synthesize a target speaker in a specified acoustic environment. We characterize the acoustic environment by two components: additive background sound and acoustic effects associated with reverberation. Evaluating EnvTTS requires assessing both speaker preservation and environment reproduction, yet systematic evaluation remains challenging due to the lack of a dedicated benchmark and metrics that separately characterize background and reverberation similarity. To address these gaps, we introduce **EnvTTS-Bench**, a benchmark of 3,000 evaluation instances with independently specified speaker and environment references drawn from real recordings, and **EnvSim**, an environment-similarity metric that separately measures background and reverberation similarity. We also establish two baselines: a training-free cascaded approach and **EnvTTS-DiT**, a flow-matching TTS model conditioned on separate speaker and environment references. Objective and human evaluations show that EnvSim aligns strongly with human judgments for separately measuring background and reverberation similarity. System evaluations on EnvTTS-Bench further show that the cascaded approach provides a practical baseline for EnvTTS, while EnvTTS-DiT achieves stronger overall performance; together, they provide useful reference systems for future EnvTTS research.

## 1 INTRODUCTION

Recent advances in zero-shot text-to-speech (TTS) have substantially improved the ability to clone unseen speakers from reference recordings (Chen et al., 2025; Anastassiou et al., 2024; Peng et al., 2024; Hu et al., 2026). Some systems (Hai et al., 2026; Xin et al., 2026; Zhou et al., 2026) can further preserve other acoustic attributes present in the reference, including characteristics of the recording environment. However, zero-shot TTS typically derives both speaker identity and recording environment from the same reference, making them difficult to control independently. Enabling separate control of speaker and environment would benefit content creation and speech data augmentation. For example, content creation may require dubbed speech to match the acoustic environment of the original recording, while data augmentation can generate speech from the same speaker in different environments for training ASR and speaker verification systems.

These needs motivate **environment-controllable TTS (EnvTTS)** (Tan et al., 2022; Lu et al., 2025; 2026; Glazer et al., 2025), which extends zero-shot TTS with separate speaker and environment references. Given a speaker reference, an environment reference, and target text, EnvTTS aims to generate speech with the speaker identity specified by the speaker reference and the acoustic characteristics specified by the environment reference. Unlike prior studies that focus primarily on background conditions, our formulation of EnvTTS considers both additive background sound and acoustic transformations applied to speech, such as reverberation.

Existing evaluations of EnvTTS (Lu et al., 2025; 2026; Glazer et al., 2025) are typically conducted on limited sets of acoustic conditions and often rely on simulated environments, making it difficult to assess performance under the diversity of real recordings. Moreover, the lack of a common bench-



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.

mark hinders direct comparison across systems. We therefore introduce **EnvTTS-Bench**, a benchmark of 3,000 evaluation instances constructed from four real-world meeting corpora: AMI (McCowan et al., 2005), AliMeeting (Yu et al., 2022), ICSI (Janin et al., 2003), and NOTSOFAR-1 (Vinnikov et al., 2024). The benchmark covers diverse speakers, recording devices, rooms, and acoustic conditions, providing a common basis for evaluating speaker preservation and environment reproduction under realistic recordings.

Evaluating EnvTTS requires measuring both speaker preservation and environment reproduction. Speaker preservation can be assessed with established speaker-similarity models (Desplanques et al., 2020; Chung et al., 2020; Yakovlev et al., 2024) that are robust to recording conditions. Environment reproduction is harder to evaluate, especially in target-free settings where the generated speech must be compared directly with an environment reference containing different speaker and linguistic content. Existing representations may encode such irrelevant information (Wu et al., 2023) and often do not distinguish background noise from reverberation (Dinkel et al., 2024; Tjandra et al., 2025; Hines et al., 2015). Because background noise is additive whereas reverberation modifies speech through the acoustic response, a single similarity score can obscure which component is reproduced. We therefore introduce **EnvSim**, a learned metric that separately measures background-noise and reverberation similarity while reducing sensitivity to speaker and linguistic content.

With this evaluation framework, we benchmark a training-free cascaded EnvTTS system and an end-to-end EnvTTS model. The cascaded system combines the speaker reference with background sound extracted from the environment reference to form an environment-conditioned prompt for a frozen zero-shot TTS model. For the end-to-end approach, we introduce **EnvTTS-DiT**, a flow-matching model that directly conditions generation on separate speaker and environment references without explicit environment decomposition.

We conduct comprehensive evaluations of EnvSim and the proposed EnvTTS systems on EnvTTS-Bench. Objective and human evaluations show that EnvSim effectively separates background-noise and reverberation similarity and aligns well with human perception. System evaluations further show that the cascaded approach provides a reasonable baseline for EnvTTS, whereas EnvTTS-DiT achieves stronger overall performance across environment fidelity, speaker similarity, and intelligibility, establishing a strong end-to-end reference system for future EnvTTS research.

In summary, we provide a comprehensive foundation for EnvTTS[1] by introducing **EnvTTS-Bench**, a benchmark of 3,000 evaluation instances drawn from four real-world meeting corpora with diverse speakers and acoustic environments, and **EnvSim**, an environment-similarity metric that separately assesses background-noise and reverberation similarity. We further develop two EnvTTS systems: a training-free cascaded baseline and **EnvTTS-DiT**, an end-to-end model conditioned on separate speaker and environment references, providing practical baselines for future EnvTTS research.

## 2 Task Formulation and Evaluation

### 2.1 EnvTTS Definition

Given target text $c$, a speaker reference $x_{\text{spk}}$, and an environment reference $x_{\text{env}}$, EnvTTS generates

$$\hat{x} = G(c, x_{\text{spk}}, x_{\text{env}}), \tag{1}$$

where $\hat{x}$ should follow the linguistic content of $c$, preserve the speaker identity of $x_{\text{spk}}$, and reproduce the acoustic conditions of $x_{\text{env}}$. In other words, EnvTTS synthesizes how the target speaker would sound when speaking the target text in the specified environment.

We characterize an acoustic environment by two components: acoustic transformations of speech and additive background sound. A recording can be approximated as

$$x(t) = s(t) * h(t) + n(t), \tag{2}$$

where $s(t)$ is the underlying clean speech, $h(t)$ represents the acoustic response of the environment, primarily characterized by the room impulse response (RIR), $n(t)$ is the additive background sound, and $*$ denotes convolution. The acoustic response captures effects such as propagation and reverberation, while the background component is added to the resulting signal. Environment reproduction therefore requires matching both components.

[1] Anonymous project page: `https://anonymous2027iclr.github.io/EnvTTS/`.

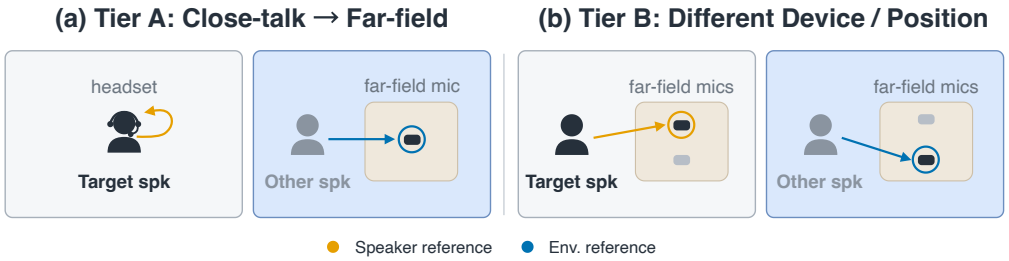


Figure 1: Overview of EnvTTS-Bench.

Table 1: Statistics of EnvTTS-Bench. Duration$_T$ denotes the mean duration of the target recording.

| Corpus | Language | # Sessions | # Triplets | Duration$_T$ |
|---|---|---|---|---|
| *Tier A: close-talk speaker reference* | | | | |
| AMI | EN | 16 | 738 | 4.9 s |
| AliMeeting | ZH | 20 | 762 | 5.4 s |
| *Tier B: far-field cross-device speaker reference* | | | | |
| ICSI | EN | 75 | 1,090 | 5.9 s |
| NOTSOFAR-1 | EN | 72 | 410 | 6.4 s |
| **Total** | — | 183 | 3,000 | 5.6 s |

### 2.2 EnvTTS-Bench

**Overview.** EnvTTS-Bench is designed to evaluate environment-controllable TTS under diverse real recording conditions. We construct **EnvTTS-Bench** with independently specified speaker and environment references drawn from real recordings. Each instance contains a speaker reference $x_{\text{spk}}$ and an environment reference $x_{\text{env}}$, where the two recordings contain different speakers and are recorded under different acoustic conditions. The speaker reference specifies the target speaker identity, while the environment reference specifies the acoustic environment to be reproduced. By separating the two references, the benchmark requires systems to preserve speaker identity while adapting to an independently specified recording environment.

**Benchmark construction.** We construct EnvTTS-Bench from four multi-channel meeting corpora: AMI (McCowan et al., 2005), AliMeeting (Yu et al., 2022), ICSI (Janin et al., 2003), and NOTSOFAR-1 (Vinnikov et al., 2024). These corpora span different rooms, microphone positions, array configurations, and recording devices, providing substantial variation in background conditions and acoustic response. For each instance, the speaker and environment references are selected independently from different speakers, preventing the environment reference from directly revealing the target speaker identity.

Figure 1 illustrates the two-tier construction of EnvTTS-Bench. We divide the benchmark by the recording condition of the speaker reference. **Tier A** includes AMI and AliMeeting, where the speaker reference comes from a close-talk channel and the environment reference from a far-field recording. **Tier B** includes ICSI and NOTSOFAR-1, where we deliberately draw the speaker reference from a mismatched far-field channel rather than the close-talk channel. In ICSI, the mismatch comes from a different microphone position, while in NOTSOFAR-1 it also involves a different recording device. Tier B therefore introduces a stronger acoustic mismatch between the speaker and environment references.

Candidate references are filtered for unsuitable acoustic conditions and annotation or segmentation errors, followed by sanity checks and manual listening. To diversify background conditions, half of the environment references are augmented with MUSAN noise (Snyder et al., 2015) at controlled SNRs. Detailed filtering, sampling, and augmentation procedures are provided in Appendix B. As a result, EnvTTS-Bench contains 3,000 instances from 183 sessions across four corpora, evenly split between the two tiers and covering both English and Mandarin, as summarized in Table 1.

### 2.3 ENVSIM

Motivated by the acoustic decomposition in Equation 2, evaluating environment reproduction requires measuring both RIR-induced acoustic transformation and additive background sound, while remaining insensitive to speaker identity and linguistic content. We therefore introduce **EnvSim**, an embedding model that focuses on background-noise and reverberation characteristics for environment similarity assessment.

#### 2.3.1 ENVSIM DESIGN

Given an input recording $x$, a mel-spectrogram-based audio encoder $g(\cdot)$ (Dinkel et al., 2024) produces frame-level representations

$$H = g(x). \tag{3}$$

Two separate prediction heads then produce background-noise and reverberation embeddings,

$$z_{\text{noise}} = h_{\text{noise}}(H), \qquad z_{\text{rir}} = h_{\text{rir}}(H). \tag{4}$$

Each head uses attentive statistics pooling followed by a linear projection, and the resulting embeddings are $\ell_2$-normalized.

Given two recordings $x_i$ and $x_j$, EnvSim computes their background-noise and reverberation similarities as

$$S_{\text{noise}}(x_i, x_j) = \left(z_{\text{noise}}^{(i)}\right)^\top z_{\text{noise}}^{(j)}, \qquad S_{\text{rir}}(x_i, x_j) = \left(z_{\text{rir}}^{(i)}\right)^\top z_{\text{rir}}^{(j)}. \tag{5}$$

#### 2.3.2 ENVSIM TRAINING

**Factorial data construction.** We construct training recordings by applying independently sampled background noise and RIRs to clean speech. Within each training group, $N$ noise instances and $N$ RIRs are combined into an $N \times N$ factorial grid, with different clean speech segments assigned across the resulting combinations to vary speaker identity and linguistic content. This construction independently varies noise, reverberation, speaker, and content, so positive pairs are matched only by the target environmental factor. Details of the speech, noise, and RIR sources and the sampling procedure are provided in Appendix C.1.

**Contrastive objective.** As illustrated in Figure 2 (a), EnvSim is trained with separate contrastive objectives for background noise and reverberation. For each factor $c \in \{\text{noise}, \text{rir}\}$, positives share the same value of $c$ while differing in the other factor $c'$, whereas negatives differ in $c$. For recording $i$ with labels $(y_i^{\text{noise}}, y_i^{\text{rir}})$, we define

$$P_i^c = \left\{ j : y_j^c = y_i^c,\ y_j^{c'} \neq y_i^{c'} \right\}, \qquad N_i^c = \left\{ j : y_j^c \neq y_i^c \right\}. \tag{6}$$

Accordingly, the noise branch treats recordings with the same noise but different RIRs as positives, while the reverberation branch treats recordings with the same RIR but different noise as positives. Each branch is trained with an InfoNCE objective (van den Oord et al., 2018; Radford et al., 2021),

$$\mathcal{L}_c = -\frac{1}{B} \sum_{i=1}^{B} \frac{1}{|P_i^c|} \sum_{p \in P_i^c} \log \frac{\exp\left((z_i^c)^\top z_p^c / \tau\right)}{\sum_{a \in P_i^c \cup N_i^c} \exp\left((z_i^c)^\top z_a^c / \tau\right)}. \tag{7}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{rir}}. \tag{8}$$

This factorized training encourages each branch to retain its target environmental factor while remaining invariant to the other.

## 3 ENVTTS SYSTEMS

We consider systems that take target text $c$, a speaker reference $x_{\text{spk}}$, and an environment reference $x_{\text{env}}$ as input, and generate speech $\hat{x}$ that combines the target content, speaker identity, and acoustic environment defined in Section 2.1. We study both a cascaded baseline that transfers background conditions through prompt construction and an end-to-end model that directly conditions generation on the two references.

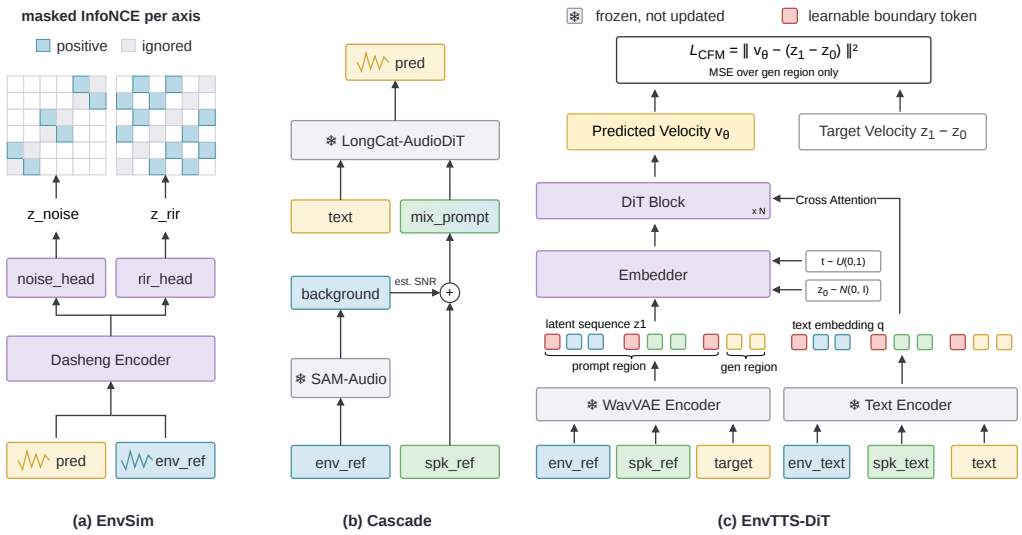


Figure 2: Overview of EnvSim and EnvTTS systems.

### 3.1 Cascaded Baseline

The cascaded baseline is a training-free pipeline that constructs an environment-conditioned speaker prompt for a pretrained zero-shot TTS model. As illustrated in Figure 2 (b), a pretrained source-separation model (Shi et al., 2025) first extracts the background component from the environment reference $x_{\text{env}}$. The extracted background is then mixed with the speaker reference $x_{\text{spk}}$ according to the estimated SNR of $x_{\text{env}}$, producing a prompt that combines the target speaker identity with the desired background condition. Together with the target text, this prompt is fed into a frozen zero-shot TTS model (Xin et al., 2026) to generate the output speech. The pipeline requires no additional training and serves as a simple baseline for transferring background conditions through prompt construction. Further details are provided in Appendix C.3.

### 3.2 EnvTTS-DiT

EnvTTS-DiT extends a pretrained conditional flow-matching TTS model to support independent speaker and environment conditioning. Rather than explicitly extracting environmental components, it directly conditions generation on the target text, speaker reference, and unseparated environment reference, as illustrated in Figure 2 (c).

**Backbone.** We build EnvTTS-DiT on LongCat AudioDiT (Xin et al., 2026). A frozen WavVAE encodes 24 kHz audio into latent representations, while a frozen UMT5 encoder maps text into text embeddings:

$$z_1 = \text{WavVAE}(x), \qquad q = \text{UMT5}(c). \tag{9}$$

Given Gaussian noise $z_0 \sim \mathcal{N}(0, I)$, flow matching constructs

$$z_t = (1 - t)z_0 + tz_1, \qquad t \sim \mathcal{U}(0, 1), \tag{10}$$

and trains the DiT to predict the corresponding velocity field. We retain the pretrained WavVAE and text encoder and adapt the DiT backbone for multi-reference conditioning.

**Multi-reference conditioning.** Standard voice-cloning prompting uses a single reference and does not distinguish speaker from environment conditioning. We instead organize the input into environment, speaker, and target streams, marked by learned boundary tokens BOE, BOS, and BOT, respectively. The audio latent sequence is

$$z_1 = \left[\langle \text{BOE} \rangle^z \| z_{\text{env}} \| \langle \text{BOS} \rangle^z \| z_{\text{spk}} \| \langle \text{BOT} \rangle^z \| z_{\text{tgt}}\right], \tag{11}$$

Table 2: Evaluation of environment embeddings on background and RIR similarity.

| Method | Background | | | | | RIR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | EER ↓ | AUC ↑ | mAP ↑ | P@1 ↑ | R@5 ↑ | EER ↓ | AUC ↑ | mAP ↑ | P@1 ↑ | R@5 ↑ |
| Random | 0.500 | 0.500 | 0.015 | 0.015 | 0.073 | 0.500 | 0.500 | 0.014 | 0.014 | 0.069 |
| Dasheng | 0.217 | 0.864 | 0.412 | 0.699 | 0.837 | 0.432 | 0.588 | 0.047 | 0.116 | 0.300 |
| CLAP | 0.187 | 0.889 | 0.510 | 0.865 | 0.957 | 0.480 | 0.519 | 0.025 | 0.041 | 0.134 |
| AES-PQ | 0.405 | 0.630 | 0.042 | 0.068 | 0.230 | 0.475 | 0.538 | 0.025 | 0.039 | 0.137 |
| EnvSim | **0.022** | **0.995** | **0.968** | **0.997** | **0.999** | **0.096** | **0.963** | **0.732** | **0.887** | **0.967** |
| opp. head | 0.510 | 0.478 | 0.023 | 0.063 | 0.143 | 0.521 | 0.460 | 0.017 | 0.020 | 0.062 |

and the text-conditioning sequence follows the same layout:

$$q = \left[\langle \mathrm{BOE} \rangle^q \| q_{\mathrm{env}} \| \langle \mathrm{BOS} \rangle^q \| q_{\mathrm{spk}} \| \langle \mathrm{BOT} \rangle^q \| q_{\mathrm{tgt}}\right]. \tag{12}$$

The audio- and text-space boundary tokens are learned separately, and each reference is paired with its transcript. This allows the model to condition directly on the full acoustic information in the environment reference rather than on an explicitly decomposed or discretized environment representation.

**Training.** We construct training examples with separate environment, speaker, and target recordings following the multi-reference formulation above. Details of the training corpora, triplet construction, and sampling strategy are provided in Appendix C.2.

During training, the environment and speaker streams are treated as context, while the flow-matching objective is applied only to the target region. Let $m$ denote a binary mask over the prompt region. The context latent is

$$z_{\mathrm{ctx}} = m \odot z_1, \tag{13}$$

and the masked conditional flow-matching objective is

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}_{t,z_0,z_1}\left[\|(1-m) \odot ((z_1 - z_0) - v_\theta(z_t, t, z_{\mathrm{ctx}}, q))\|_2^2\right]. \tag{14}$$

Thus, both references remain available as conditioning context throughout training but are excluded from the regression target.

**Inference.** At inference time, the environment reference, speaker reference, and target text are arranged using the same three-stream layout. The target audio region is initialized from noise and generated through the learned flow, while the speaker and environment streams remain fixed as conditioning context. EnvTTS-DiT therefore conditions directly on the original environment recording without requiring explicit noise extraction or RIR estimation.

# 4 Experiments

## 4.1 EnvSim Evaluation

**Setup.** We compare EnvSim against three pretrained audio representations. **Dasheng** (Dinkel et al., 2024) uses the same pretrained encoder as EnvSim but without factor-specific fine-tuning or projection heads, with cosine similarity computed between mean-pooled embeddings. **CLAP** (Wu et al., 2023) uses cosine similarity between pretrained audio embeddings. **AES-PQ**, the Production Quality score from Audiobox Aesthetics (Tjandra et al., 2025), provides a scalar estimate of acoustic quality and serves as a non-factorized acoustic baseline. EnvSim uses Dasheng as its shared audio encoder. Training details are provided in Appendix C.1.

**Objective evaluation.** We evaluate EnvSim on two environment-similarity tasks: background similarity and RIR similarity. For background similarity, two recordings are considered matched if they share the same background sound regardless of their RIRs; for RIR similarity, they are considered matched if they share the same RIR regardless of their background sounds. Each task is evaluated under both verification and retrieval protocols. Verification determines whether a pair of

Table 3: Agreement of environment similarity models with human judgments.

| Metric | $\rho_{\text{Bkg}}$ | $\rho_{\text{RIR}}$ | $\rho^{\text{P}}_{\text{Bkg}\|\text{RIR}}$ | $\rho^{\text{P}}_{\text{RIR}\|\text{Bkg}}$ |
|---|---|---|---|---|
| Dasheng | 0.507 | 0.453 | 0.441 | 0.373 |
| CLAP | 0.474 | 0.383 | 0.412 | 0.294 |
| AES-PQ | 0.511 | 0.297 | 0.466 | 0.184 |
| EnvSim | **0.621** | **0.639** | **0.601** | **0.661** |
| opp. head | 0.021 | 0.201 | −0.219 | 0.031 |

recordings is matched along the evaluated axis, and we report equal error rate (EER) and area under the ROC curve (AUC); retrieval uses each recording as a query against the remaining recordings and reports mean average precision (mAP), precision at rank 1 (P@1), and recall at rank 5 (R@5).

The evaluation set contains 2,304 held-out recordings organized into 16 factorial groups, each covering all $12 \times 12$ combinations of 12 background instances and 12 RIRs with randomly sampled speech. All speech, background, and RIR sources are disjoint from those used for training. Details of the evaluation-set construction and protocol are provided in Appendix C.1.

As shown in Table 2, the general-purpose baselines are more sensitive to background similarity than to RIR similarity. Among them, AES-PQ performs substantially worse on both tasks, consistent with its design as an overall audio-quality predictor rather than an environment-similarity representation. CLAP performs better on background similarity, while Dasheng is comparatively stronger on RIR similarity. In contrast, EnvSim achieves the best performance on both factors, indicating its effectiveness in separately modeling background and reverberation similarity.

We further examine axis selectivity using the *opposite head* setting, where each branch is evaluated on the other environmental axis. Its EER is near chance and retrieval performance is close to random, in contrast to the strong performance on the corresponding axis. This separation indicates that the two branches learn distinct representations of background and RIR similarity.

**Agreement with human judgments.** We construct 60 recording pairs that independently vary whether the background and RIR are matched, spanning real EnvTTS-Bench recordings, synthetic acoustic renders, EnvTTS-DiT outputs, and CASCADE outputs to cover diverse acoustic and generation conditions. Speaker identity and linguistic content are allowed to vary across pairs. A total of 15 listeners with backgrounds in audio engineering or music production rate each pair on separate 1–5 scales for background and reverberation similarity. Further details of the subjective evaluation are provided in Appendix A.1.

Table 3 reports Spearman rank correlation (Spearman, 1904), denoted by $\rho$, between metric scores and mean listener ratings. EnvSim uses the corresponding branch for each similarity axis, whereas each single-score baseline uses the same score for both. EnvSim achieves the strongest correlation on both background and reverberation similarity, and its improvement over Dasheng, which shares the same pretrained encoder, highlights the benefit of learning separate environment representations.

Because background and reverberation judgments may themselves be correlated, we additionally report partial Spearman correlation (Kendall, 1942), denoted by $\rho^{\text{P}}$, for each perceptual axis while controlling for the other. EnvSim remains strongly correlated with human judgments on the corresponding axis, while the *opposite head* shows little correlation. Together, these results show that EnvSim aligns with human perception while maintaining separation between background and reverberation similarity.

### 4.2 ENVTTS EVALUATION

**Setup.** We evaluate EnvTTS-DiT against CASCADE on EnvTTS-Bench. CASCADE uses SAM-Audio (Shi et al., 2025) to extract background sound from the environment reference, mixes it with the speaker reference at the estimated SNR, and uses the resulting mixture as the TTS prompt. We also include two single-reference anchors: CLONE-ENV, which conditions only on the environment reference, and CLONE-SPK, which conditions only on the speaker reference. These anchors charac-

Table 4: Objective and subjective results on EnvTTS-Bench.

| System | WER ↓ | $\text{EnvSim}_{\text{Avg}}$ ↑ | $\text{EnvSim}_{\text{Bkg}}$ ↑ | $\text{EnvSim}_{\text{RIR}}$ ↑ | SpkSim ↑ | $\text{MOS}_{\text{Env}}$ ↑ | $\text{MOS}_{\text{Spk}}$ ↑ | $\text{MOS}_{\text{Nat}}$ ↑ |
|---|---|---|---|---|---|---|---|---|
| CLONE-ENV | 13.19 | 0.660 | 0.692 | 0.627 | 0.115 | — | — | — |
| CLONE-SPK | 11.73 | 0.360 | 0.453 | 0.267 | 0.683 | — | — | — |
| CASCADE | 13.70 | 0.414 | 0.566 | 0.263 | 0.475 | $2.82 \pm 0.11$ | $3.11 \pm 0.07$ | $3.12 \pm 0.07$ |
| EnvTTS-DiT | **8.38** | **0.562** | **0.690** | **0.434** | **0.480** | $\mathbf{3.77} \pm 0.09$ | $\mathbf{3.34} \pm 0.07$ | $\mathbf{3.18} \pm 0.10$ |

Table 5: Ablation of the EnvSim encoder backbone.

| Encoder | Background | | | | | RIR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | EER ↓ | AUC ↑ | mAP ↑ | P@1 ↑ | R@5 ↑ | EER ↓ | AUC ↑ | mAP ↑ | P@1 ↑ | R@5 ↑ |
| BEATs | **0.017** | **0.996** | **0.971** | **0.998** | **0.999** | **0.095** | **0.963** | 0.661 | 0.818 | 0.954 |
| Dasheng | 0.022 | 0.995 | 0.968 | 0.997 | **0.999** | 0.096 | **0.963** | **0.732** | **0.887** | **0.967** |

terize single-reference conditioning rather than complete EnvTTS systems. Training and inference details for all systems are provided in Appendix C.

For objective evaluation, we measure linguistic content using word error rate (WER) with Qwen3-ASR (Shi et al., 2026), speaker similarity using ReDimNet-B6 (Yakovlev et al., 2024)[2], and environment similarity using EnvSim. Speaker similarity is computed against the speaker reference, while environment similarity is computed against the environment reference. We report $\text{EnvSim}_{\text{Bkg}}$, $\text{EnvSim}_{\text{RIR}}$, and their average $\text{EnvSim}_{\text{Avg}}$. For subjective evaluation, 15 listeners rate environment similarity ($\text{MOS}_{\text{Env}}$), speaker similarity ($\text{MOS}_{\text{Spk}}$), and naturalness ($\text{MOS}_{\text{Nat}}$) on 1–5 scales for 30 instances sampled from both tiers of EnvTTS-Bench. EnvTTS-DiT and CASCADE use the same text and references for each instance, with system identities hidden during evaluation. Further details of the subjective evaluation are provided in Appendix A.2.

**Results.** As shown in Table 4, the single-reference anchors exhibit the expected trade-off: CLONE-ENV achieves high environment similarity but low speaker similarity, while CLONE-SPK shows the opposite pattern. These two anchors provide approximate upper bounds for environment and speaker preservation under single-reference conditioning, but neither performs the full EnvTTS task. CASCADE preserves speaker identity while transferring background characteristics, but provides limited RIR similarity because it explicitly transfers only the extracted background sound. In contrast, EnvTTS-DiT improves both background and RIR similarity over CASCADE, while achieving lower WER and comparable speaker similarity. This suggests that direct conditioning on the environment reference enables EnvTTS-DiT to capture acoustic characteristics beyond additive background sound, including reverberation to some extent.

The subjective results show the same overall trend. EnvTTS-DiT substantially improves perceived environment similarity over CASCADE, while achieving slightly higher speaker similarity and comparable naturalness. Overall, EnvTTS-DiT provides stronger joint control of speaker identity and acoustic environment than the cascaded baseline.

### 4.3 ABLATIONS

**EnvSim backbone.** We compare Dasheng and BEATs (Chen et al., 2023) as the shared audio encoder under the same EnvSim training objective. As shown in Table 5, the two encoders perform similarly on verification and background retrieval, while Dasheng performs better on RIR retrieval. One possible reason is that BEATs operates on patch-level representations with lower temporal resolution, which may be less suited to capturing fine-grained reverberation cues. We therefore use Dasheng as the EnvSim encoder.

**EnvTTS-DiT stream design.** We ablate the boundary tokens and stream ordering. We compare the full environment-first configuration $[\text{env}, \text{spk}, \text{tgt}]$ with a variant without boundary tokens and a speaker-first variant $[\text{spk}, \text{env}, \text{tgt}]$. All variants are trained for 1,000 steps under identical settings.

[2]We use the official `b6-vox2-ft_lm.pt` checkpoint.

Table 6: Ablation of the EnvTTS-DiT stream design at 1,000 training steps.

| Variant | WER ↓ | $\text{EnvSim}_{\text{Avg}}$ ↑ | $\text{EnvSim}_{\text{Bkg}}$ ↑ | $\text{EnvSim}_{\text{RIR}}$ ↑ | SpkSim ↑ |
|---|---|---|---|---|---|
| EnvTTS-DiT | 9.24 | **0.571** | 0.675 | **0.467** | 0.444 |
| w/o bnd. tokens | 9.91 | 0.570 | **0.685** | 0.456 | **0.470** |
| Speaker-first | **9.06** | 0.558 | 0.682 | 0.434 | 0.460 |

As shown in Table 6, removing the boundary tokens has little effect on average environment similarity but reduces RIR similarity and increases WER. The speaker-first ordering further lowers both average and RIR similarity, despite small gains in WER, background similarity, and speaker similarity. We therefore retain the boundary tokens and environment-first ordering for better environment preservation, particularly reverberation.

## 5 Related Work

**Environment Control in Text-to-Speech.** Prior work on environment-controllable TTS can be broadly divided by whether speaker identity is independently controlled. Some methods control the acoustic environment through reference audio, learned environment representations, or textual prompts, while speaker identity is fixed or not separately specified (Tan et al., 2022; Lee et al., 2024; Jung et al., 2025; Yun et al., 2026; Wang et al., 2026). Other methods enable separate control of speaker and environment references (Lu et al., 2025; 2026; Glazer et al., 2025). However, these studies primarily focus on background conditions, with acoustic effects such as reverberation receiving less systematic treatment. Evaluation also lacks a common benchmark covering diverse real recording conditions, and released checkpoints are limited, making direct comparison difficult. Our work addresses these gaps with a real-recording benchmark, explicit evaluation of both background and reverberation, and reproducible EnvTTS baselines.

**Environment Similarity Metrics.** Reference-based metrics such as PESQ (Rix et al., 2001) and ViSQOL (Hines et al., 2015) generally assume aligned or matched speech and are not designed to characterize acoustic environments. General-purpose embeddings such as CLAP (Wu et al., 2023) can capture diverse acoustic conditions, but may jointly encode content, speaker, and acoustic scene. Meta Audiobox Aesthetics (Tjandra et al., 2025) predicts overall audio quality without explicitly separating background noise from reverberation. In contrast, EnvSim is designed specifically for environment similarity and separately measures background and reverberation.

## 6 Conclusion

We studied EnvTTS, a text-to-speech setting that enables independent control of speaker identity and acoustic environment. To support reliable evaluation, we presented EnvTTS-Bench, a benchmark of 3,000 evaluation instances with independently specified speaker and environment references drawn from real recordings, and EnvSim, an environment-similarity metric that separately measures background and RIR similarity. We further studied two EnvTTS systems: a training-free cascaded approach and EnvTTS-DiT, an end-to-end flow-matching model with separate speaker and environment conditioning. Objective and subjective evaluations show that EnvSim aligns well with human judgments of background and reverberation similarity, while the cascaded approach and EnvTTS-DiT provide practical reference systems for environment-controllable TTS.

In future work, we will explore incorporating more explicit acoustic or spatial cues to improve reverberation modeling, particularly when RIR-related information is difficult to infer reliably from short environment references. We also plan to extend EnvTTS-Bench and model training beyond meeting-style recordings to a broader range of acoustic environments, devices, and recording configurations, enabling more comprehensive evaluation under real-world conditions.

### AI USE STATEMENT

We used generative AI tools during the preparation of this manuscript. These tools were used primarily to assist with language polishing, improving clarity and organization of the writing, and supporting literature retrieval and discovery of potentially relevant prior work. All retrieved references were manually verified by the authors before inclusion. The AI tools were not used to generate experimental results, conduct data analysis, make scientific claims, or determine the conclusions of the paper. The authors take full responsibility for the content, correctness, citations, and conclusions presented in this work.

### ETHICS STATEMENT

This work studies environment-controllable text-to-speech and involves the synthesis and evaluation of speech recordings. All datasets used in this work are publicly available research corpora and are used in accordance with their respective licenses and intended research purposes. Participants in the subjective listening studies provided informed consent before taking part, and no sensitive personal information was collected.

### REPRODUCIBILITY STATEMENT

We provide detailed descriptions of EnvTTS-Bench construction, EnvSim training and evaluation, the cascaded EnvTTS baseline, EnvTTS-DiT architecture and training, inference settings, and subjective evaluation protocols in the main paper and appendix. We also report the datasets, pretrained TTS and audio models, checkpoints, and evaluation metrics used in our experiments. Code, benchmark metadata, evaluation scripts, and model checkpoints will be released through the project page to support reproducibility and future comparison.

## REFERENCES

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193, 2023.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025.

Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee-Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. In defence of metric learning for speaker recognition. In *Proceedings of Interspeech*, pp. 2977–2981, 2020. doi: 10.21437/Interspeech.2020-1064.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proceedings of Interspeech*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang. Scaling up masked audio encoder learning for general audio classification. In *Interspeech 2024*, pp. 547–551, 2024. doi: 10.21437/Interspeech.2024-246.

Neta Glazer, Aviv Navon, Yael Segal, Aviv Shamsian, Hilit Segev, Asaf Buchnick, Menachem Pirchi, Gil Hetz, and Joseph Keshet. Umbratts: Adapting text-to-speech to environmental contexts with flow matching. *arXiv preprint arXiv:2506.09874*, 2025.

Jiarui Hai, Karan Thakkar, Ke Chen, Yunyun Wang, Jiaqi Su, Rithesh Kumar, Mounya Elhilali, and Zeyu Jin. Voicedesigner: Text-to-voice generation and editing via unified diffusion modeling and data augmentation. *arXiv preprint arXiv:2608.13613*, 2026.

Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Adam Janin, Don Baron, Jane Edwards, Daniel P. W. Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pp. I–364–I–367, 2003. doi: 10.1109/ICASSP.2003.1198793.

Jaemin Jung, Junseok Ahn, Chaeyoung Jung, Tan Dat Nguyen, Youngjoon Jang, and Joon Son Chung. Voicedit: Dual-condition diffusion transformer for environment-aware speech synthesis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Maurice G. Kendall. Partial rank correlation. *Biometrika*, 32(3/4):277–283, 1942.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5220–5224. IEEE, 2017.

Yeonghyeon Lee, Inmo Yeon, Juhan Nam, and Joon Son Chung. Voiceldm: Text-to-speech with environmental context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12566–12571. IEEE, 2024.

Ye-Xin Lu, Hui-Peng Du, Zheng-Yan Sheng, Yang Ai, and Zhen-Hua Ling. Incremental disentanglement for environment-aware zero-shot text-to-speech synthesis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Ye-Xin Lu, Yu Gu, Kun Wei, Hui-Peng Du, Yang Ai, and Zhen-Hua Ling. Daien-tts: Disentangled audio infilling for environment-aware text-to-speech synthesis. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17287–17291. IEEE, 2026.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI meeting corpus. In *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, pp. 137–140, 2005.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12442–12462, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus, 2015. URL `https://arxiv.org/abs/1510.08484`.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Daxin Tan, Guangyan Zhang, and Tan Lee. Environment Aware Text-to-Speech Synthesis. In *Interspeech 2022*, pp. 481–485, 2022. doi: 10.21437/Interspeech.2022-10348.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Alon Vinnikov, Amir Ivry, Aviv Hurvitz, Igor Abramovski, Sharon Koubi, Ilya Gurvich, Shai Peer, Xiong Xiao, Benjamin Martinez Elizalde, Naoyuki Kanda, Xiaofei Wang, Shalev Shaer, Stav Yagev, Yossi Asher, Sunit Sivasankaran, Yifan Gong, Min Tang, Huaming Wang, and Eyal Krupka. NOTSOFAR-1 challenge: New datasets, baseline, and tasks for distant meeting transcription. In *Proceedings of Interspeech*, pp. 5003–5007, 2024. doi: 10.21437/Interspeech.2024-1788.

Helin Wang, Jiarui Hai, Dading Chong, Karan Thakkar, Tiantian Feng, Dongchao Yang, Junhyeok Lee, Thomas Thebaud, Laureano Moro-Velazquez, Jesus Villalba, Zengyi Qin, Shrikanth Narayanan, Mounya Elhilali, and Najim Dehak. Capspeech: Enabling downstream applications in style-captioned text-to-speech. *IEEE Transactions on Audio, Speech and Language Processing*, 34:3519–3531, 2026. doi: 10.1109/TASLPRO.2026.3703247.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Detai Xin, Shujie Hu, Chengzuo Yang, Chen Huang, Guoqiao Yu, Guanglu Wan, and Xunliang Cai. Longcat-audiodit: High-fidelity diffusion text-to-speech in the waveform latent space. *arXiv preprint arXiv:2603.29339*, 2026.

Ivan Yakovlev, Rostislav Makarov, Andrei Balykin, Pavel Malov, Anton Okhotnikov, and Nikita Torgashov. Reshape dimensions network for speaker recognition. In *Proceedings of Interspeech*, pp. 3235–3239, 2024. doi: 10.21437/Interspeech.2024-2116.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Jun-Hak Yun, Seung-Bin Kim, and Seong-Whan Lee. Immersivetts: Environment-aware text-to-speech with multimodal diffusion transformer and domain-specific representation alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38295–38314, 2026.

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. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, pp. 35139–35148, 2026.

# A SUBJECTIVE EVALUATION PROTOCOL

Fifteen raters participate in all listening tests, and each rater evaluates all trials in each test. To limit listener fatigue, the trials of each test are divided into two non-overlapping halves, A and B, each taking approximately 20 minutes to complete. The two halves are completed in separate sittings with a break in between. Both tests are administered through a self-contained browser interface. Raters enter an assigned rater ID (1–15), select half A or B, and complete one trial per page. They may replay each clip and revisit earlier trials, and their responses can be exported as a CSV file at the end of each session. Raters are instructed to use wired headphones in a quiet room and to keep the playback volume fixed after calibration.

## A.1 ENVSIM SIMILARITY RATING TEST

This test measures whether listeners can independently judge two components of an acoustic environment: background and RIR. Each trial consists of a pair of clips, A and B, containing different speakers saying different sentences. The 60 pairs are divided into halves A (pairs 1–30) and B (pairs 31–60). Trial order within each half is randomized per rater using a deterministic seed derived from the rater ID and half, making the randomization reproducible. Before the test, raters complete a four-example warm-up constructed by independently varying the noise and room response. The examples cover the full $2 \times 2$ design of noise same/different and room same/different, illustrating that the two dimensions should be judged independently.

**Instructions to Rater**

Listen to two recordings, **A** and **B**, and judge how similar their **recording environments** are. You may replay and switch between them freely.
**Preparation:**
• Wear wired headphones in a quiet room.
• Play the calibration clip and set a comfortable volume at which the background noise is clearly audible. Do not change the volume afterwards.
• Complete the four warm-up examples before the first half.
**Rating Dimensions:**
• **Background similarity**: Listen only to the background noise: its character (e.g., air conditioning, fan, babble, or clatter) and its relative level. 1 = completely different noise; 2 = somewhat similar but clearly different; 3 = unsure; 4 = very similar, possibly the same type; 5 = the same or nearly identical noise.
• **RIR similarity**: Listen only to the room acoustics: how large the space sounds, the apparent speaker–microphone distance, and the duration and character of the reverberation. 1 = completely different room acoustics; 2 = somewhat similar but clearly different; 3 = unsure; 4 = very similar room acoustics; 5 = the same or nearly identical room response.
**Instructions:**
• Judge the two dimensions independently. The noise may be identical while the room differs, or vice versa.
• Ignore the speaker and linguistic content; the two clips always contain different speakers and sentences. Do not rate overall audio quality or naturalness.
• Use the full scale. When unsure, answer 3 and go with your first impression.
**Trial:** Clip A: *[A.wav]*; Clip B: *[B.wav]*.

**Rater Response**

**Background similarity:** *[1–5, e.g., 5]*; **RIR similarity:** *[1–5, e.g., 1]*.
**Logged per trial:** play counts for clips A and B and time spent on the trial.

## A.2 ENVTTS SYSTEMS SUBJECTIVE TEST

This test compares EnvTTS with CASCADE on real samples from EnvTTS-Bench. We select 30 evaluation cases stratified by source corpus (AMI, AliMeeting, ICSI, and NOTSOFAR). For each case, both systems synthesize the same target sentence using the same environment and speaker

references. The assignment of the two systems to samples A and B is randomized independently for each trial before the study, and system identities are hidden from the raters. Each trial presents the target sentence, the two reference recordings, and the two synthesized samples. Raters evaluate each synthesized sample independently along three 1–5 scales, yielding six ratings per trial. The 30 trials are divided into two halves, A and B, with 15 trials each; since each trial contains two synthesized samples, each half contains 30 rated synthesized samples. Trial order within each half is fixed and identical across raters. Before the test, raters calibrate their headphone volume and complete three warm-up trials sampled from evaluation cases outside the scored set. The warm-up trials contain no reference answers or system labels, which preserves the blindness of the study.

**Instructions to Rater**

Each trial shows a sentence, two **reference** clips that are not rated, and two synthesized **samples**, A and B, of the same sentence. Play them as often as you like, then rate each sample on its own on three 1–5 scales.
**Preparation:**
• Wear headphones in a quiet room.
• Play the calibration clip and set a comfortable volume at which the background noise is clearly audible. Do not change the volume afterwards.
• Complete the warm-up trials before the first half.
**References:**
• **Environment reference**: another speaker recorded in the target environment.
• **Speaker reference**: the voice of the target speaker.
**Rating Dimensions:**
• **Environment similarity**: Does the sample sound as if it was recorded in the same space and under the same recording conditions as the environment reference? Listen to the reverberation, the background noise and the microphone distance. *Ignore who is speaking.*
• **Speaker similarity**: Does the sample sound like the same person as the speaker reference? Listen to timbre and voice quality. *Ignore the recording environment.*
• **Naturalness**: Does the sample sound like a real person speaking naturally? Listen for robotic artifacts, dropouts, odd prosody and mispronunciations. *Ignore whether the environment or the speaker matches.*
**Instructions:**
• The order of A and B is random in every trial, and system identity is hidden.
• Rate the two samples independently. Equal scores are allowed; this is not a forced choice.
**Scale:** 1 = very poor / completely different; 3 = fair / unsure; 5 = excellent / almost identical.
**Trial:** Sentence: *[target text]*; Environment reference: *[env_ref.wav]*; Speaker reference: *[spk_ref.wav]*; Sample A: *[A.wav]*; Sample B: *[B.wav]*.

**Rater Response**

**Sample A:** environment similarity *[1–5, e.g., 4]*; speaker similarity *[1–5, e.g., 3]*; naturalness *[1–5, e.g., 3]*.
**Sample B:** environment similarity *[1–5, e.g., 2]*; speaker similarity *[1–5, e.g., 3]*; naturalness *[1–5, e.g., 3]*.
**Logged per trial:** play counts of each of the four clips and time spent on the trial.

## B EnvTTS-Bench Construction Details

**Sampling and Filtering.** We sample triplets within each session subject to four constraints. First, $x_{\text{spk}}$ and $c$ share the same speaker, whereas $x_{\text{env}}$ is always drawn from a different speaker, with no same-speaker fallback, preventing direct leakage of the target speaker identity through the environment reference. Second, all three excerpts are separated by at least 10 s, reducing contextual leakage between either reference and the target utterance. Third, each utterance serves as the text source for at most one triplet, and each reference excerpt is reused in at most $R$ triplets. Fourth, each session contributes at most $S$ triplets and each speaker appears in at most $T$, preventing long sessions or highly active speakers from dominating the benchmark.

For each candidate target utterance, we sample $x_{\text{spk}}$ at random and examine up to $K$ randomly ordered candidates for $x_{\text{env}}$, retaining the first that passes background screening. Sessions are traversed in round-robin order using a fixed random seed.

**Noise augmentation.** Meeting rooms offer few distinct background types, so we add a controlled noise condition. With probability 0.5, we mix a MUSAN noise recording (Snyder et al., 2015) into $x_{\text{env}}$ at an SNR drawn from $\mathcal{U}(5, 15)$ dB and rescale the mixture if its peak amplitude exceeds 0.99. We leave $x_{\text{spk}}$ unchanged. The augmented version replaces the clean one, so each triplet appears in exactly one condition: the benchmark contains 1,518 clean and 1,482 noise-augmented triplets.

Table 7: Construction parameters of EnvTTS-Bench per corpus. $S$: maximum triplets per session; $T$: maximum triplets per speaker; $R$: maximum triplets per reference excerpt; $K$: environment references tried per text-source utterance.

| | **AMI** | **AliMeeting** | **ICSI** | **NOTSOFAR-1** |
|---|---|---|---|---|
| $x_{\text{env}}$ channel | array, ch. 1 | array, ch. 1 | boundary mic. $i$ | device $d$ |
| $x_{\text{spk}}$ channel | headset | headset | boundary mic. $j \neq i$ | other device model |
| Utterance duration goal | $\mathcal{U}(3, 15)$ s | TextGrid interval | $\mathcal{U}(3, 15)$ s | $\mathcal{U}(3, 8)$ s |
| $S$ | 80 | 80 | 40 | 20 |
| $T$ | 50 | 50 | 80 | 60 |
| $R$ | 8 | 8 | 12 | 12 |
| $K$ | 10 | 10 | 20 | 20 |
| Clean / noise-augmented | 381 / 357 | 375 / 387 | 554 / 536 | 208 / 202 |

# C Training and Inference Details

## C.1 EnvSim Training Details

**EnvSim data.** EnvSim requires supervision that separates background noise from room acoustics. We synthesize training recordings by independently sampling a noise instance $n$ and a room impulse response $r$ for each clean speech segment, yielding $x_{n,r}$. We draw $n$ from MUSAN (Snyder et al., 2015) and $r$ from the room impulse responses of Ko et al. (2017). For each factor, we sample 100 instances for training and a disjoint set of 100 for evaluation. We sample speech content, speaker identity, SNR, and noise crops independently, and none serve as supervision.

Each training batch forms an $N \times N$ factorial grid ($N$=12), built by drawing $N$ noise instances and $N$ RIRs from the corresponding pool, with one recording for every Noise–RIR combination. For the noise axis, positives share the same noise but use different RIRs; for the reverberation axis, positives share the same RIR but use different noises. We exclude pairs that match both factors and treat pairs that differ on the target axis as negatives, following Eq. 6. The cross-factor constraint ensures that a positive pair shares only the target environmental factor: in an $N \times N$ grid, each anchor has $N - 1$ positives and $N(N - 1)$ negatives on each axis, including $N - 1$ hard negatives that share the non-target factor.

**EnvSim training.** We train EnvSim on a single NVIDIA H800 GPU for 2,000 steps, fully fine-tuning the Dasheng encoder and the noise and reverberation projection heads. We use AdamW with learning rates of $1\text{e}{-5}$ for the backbone and $1\text{e}{-3}$ for the heads, with a 100-step linear warmup followed by cosine decay. Each step uses an effective batch of 432 clips, sampled from 3 blocks of 12 noise $\times$ 12 reverberation combinations.

## C.2 EnvTTS-DiT Training Details

**EnvTTS-DiT data.** Training data is split evenly between synthetic examples and real recordings. **For the synthetic half**, we construct examples online from a speaker-grouped LibriSpeech pool (Panayotov et al., 2015). The environment reference is always drawn from a different speaker than the target, while the speaker reference is independently re-augmented with a separately sampled room response and background noise. This design discourages the model from copying speaker

characteristics from the environment reference or acoustic conditions from the speaker reference. **For the real half**, we use AMI and AliMeeting recordings. Each target shares the same speaker as the speaker reference but is recorded under a different condition, while it shares the same session and microphone as the environment reference but contains a different speaker. This provides an approximate match in background and reverberation conditions between the target and environment reference. Training sessions are disjoint from the EnvTTS-Bench evaluation split to prevent evaluation data from leaking into training.

**EnvTTS-DiT training.** We fine-tune LongCat AudioDiT on a single NVIDIA H800 GPU with bf16 precision for 3,000 steps. WavVAE and the UMT5 text encoder are frozen, while all 24 DiT layers are adapted with rank-32 LoRA on the attention and feed-forward projections; newly introduced conditioning modules and embeddings are trained in full. We use AdamW with a learning rate of $5\mathrm{e}{-5}$, a 250-step linear warmup followed by cosine decay, and an effective batch size of 64.

**Augmentation parameters.** The scene $(h, n)$ drawn for each synthetic stream (an impulse response $h$ and a noise clip $n$) has each component present independently with probability $0.5$, mixed at an SNR sampled from $\mathcal{U}(-5, 15)$ dB; every signal is band-limited to the $8$ kHz speech bandwidth so that spectral extent alone cannot betray the environment.

### C.3 CASCADED BASELINE INFERENCE DETAILS

CASCADE requires no training and runs in three stages: background extraction, prompt mixing, and zero-shot synthesis. All systems receive the same preprocessed inputs. Following the training pipeline of EnvTTS-DiT, we resample every recording to 24 kHz, scale $x_{\mathrm{env}}$ and $x_{\mathrm{spk}}$ independently to an RMS level of $-23$ dBFS, hard-clip all waveforms to $[-0.5, 0.5]$, and apply the same text normalization to all transcripts and target texts.

**Background extraction.** We apply SAM-Audio (Shi et al., 2025) (`sam-audio-small`) to $x_{\mathrm{env}}$. SAM-Audio separates the source described by a text prompt from the remaining signal. We therefore describe the foreground speech with the prompt "*A person speaking*" and take the residual as the background estimate $\hat{n}$, while the separated target serves as the speech estimate $\hat{s}$. The input is resampled to the 48 kHz operating rate of SAM-Audio, and both outputs are resampled back to 24 kHz and trimmed to the input length. We use the default setting and disable candidate re-ranking and span prediction.

**SNR estimation and prompt mixing.** We estimate the speech-to-background ratio of the environment reference from the separated components,

$$\widehat{\mathrm{SNR}} = \mathrm{clip}\left(20 \log_{10} \frac{\mathrm{RMS}(\hat{s})}{\mathrm{RMS}(\hat{n})},\, -5,\, 20\right)\ \mathrm{dB}, \tag{15}$$

where clipping guards against degenerate separations, such as a near-silent residual or a residual that retains part of the speech. The background $\hat{n}$ is then tiled or truncated to the length of $x_{\mathrm{spk}}$ and mixed at the estimated SNR:

$$\tilde{x}_{\mathrm{spk}} = x_{\mathrm{spk}} + \frac{\mathrm{RMS}(x_{\mathrm{spk}})}{10^{\widehat{\mathrm{SNR}}/20}\,\mathrm{RMS}(\hat{n})}\,\hat{n}. \tag{16}$$

If the extracted background is effectively silent ($\mathrm{RMS}(\hat{n}) < 10^{-4}$), we use $x_{\mathrm{spk}}$ as the prompt without mixing.

**Zero-shot synthesis.** The mixed prompt $\tilde{x}_{\mathrm{spk}}$ conditions the frozen LongCat AudioDiT (Xin et al., 2026) in its standard single-reference voice-cloning mode. We truncate the prompt to its first 12 s and encode it with WavVAE as the context latent. Following the upstream convention, the text condition is the transcript of $x_{\mathrm{spk}}$ followed by the target text $c$. The generation length equals the duration of the target recording in each benchmark instance. The sampler settings match those of EnvTTS-DiT (Appendix C.4): an Euler solver on 16 time points with a classifier-free guidance scale of $w = 4$. The single-reference anchors CLONE-ENV and CLONE-SPK follow the same procedure, but they prompt the model directly with $x_{\mathrm{env}}$ or $x_{\mathrm{spk}}$ and the corresponding transcript, without extraction or mixing. When computing speaker similarity for CASCADE, we use the original $x_{\mathrm{spk}}$ as the reference instead of the mixed prompt.

### C.4 EnvTTS-DiT Inference Details

**Input assembly.** EnvTTS-DiT receives the same preprocessed inputs as CASCADE (Appendix C.3) and uses both references in full. The frozen WavVAE encodes $x_{\mathrm{env}}$ and $x_{\mathrm{spk}}$ separately into latent sequences $z_{\mathrm{env}}$ and $z_{\mathrm{spk}}$ with a hop size of 2,048 samples, about 11.7 frames per second. Following Eq. 11, we form the prompt latent $\big[\langle \mathrm{BOE}\rangle^z \| z_{\mathrm{env}} \| \langle \mathrm{BOS}\rangle^z \| z_{\mathrm{spk}} \| \langle \mathrm{BOT}\rangle^z\big]$ of length $T_{\mathrm{p}}$, which defines the context latent $z_{\mathrm{ctx}}$ and the prompt mask $m$. On the text side, the frozen UMT5 encoder separately encodes the transcripts of $x_{\mathrm{env}}$ and $x_{\mathrm{spk}}$ and the target text $c$, and the three embeddings are assembled with the text boundary tokens following Eq. 12. The target region contains $T_{\mathrm{g}}$ frames, set to the duration of the target recording rounded up to whole latent frames, and the total length $T_{\mathrm{p}} + T_{\mathrm{g}}$ is capped at the 30 s limit of the backbone.

**Sampling.** We draw $z_0 \sim \mathcal{N}(0, I)$ over all $T_{\mathrm{p}} + T_{\mathrm{g}}$ frames and integrate the learned velocity field from $t = 0$ to $t = 1$ with an Euler solver on a uniform grid of 16 time points. Before each velocity evaluation, we reset the prompt region of the current state to the interpolation path of the context, $m \odot z_t \leftarrow m \odot \big((1-t)\, z_0 + t\, z_{\mathrm{ctx}}\big)$, so that only the target region evolves freely. We apply classifier-free guidance (Ho & Salimans, 2022) with scale $w = 4$:

$$\tilde{v} = v_\theta(z_t, t, z_{\mathrm{ctx}}, q) + w\Big(v_\theta(z_t, t, z_{\mathrm{ctx}}, q) - v_\theta\big((1-m) \odot z_t, t, \mathbf{0}, \mathbf{0}\big)\Big). \tag{17}$$

The unconditional branch zeros the full text-conditioning sequence including its boundary tokens, the context latent, and the prompt region of the state, while keeping the attention masks unchanged. A single guidance term therefore strengthens the environment, speaker, and text conditions jointly. After integration, the WavVAE decoder converts the final $T_{\mathrm{g}}$ target frames into a 24 kHz waveform. In batched inference, we decode each sample at its exact length so that padding frames do not introduce decoder artifacts.