ImmersiveTTS: Environment-Aware Text-to-Speech with Multimodal Diffusion Transformer and Domain-Specific Representation Alignment

Jun-Hak Yun, Seung-Bin Kim, Seong-Whan Lee*
Department of Artificial Intelligence, Korea University, Seoul, Korea
{jh_yun, sb-kim, sw.lee}@korea.ac.kr

Abstract

Recent advancements in text-guided audio generation have yielded promising results in diverse domains, including sound effects, speech, and music. However, jointly generating speech with environmental audio remains challenging due to the inherent disparities in their acoustic patterns and temporal dynamics. We propose ImmersiveTTS, an environment-aware text-to-speech (TTS) model that generates natural speech seamlessly integrated within environmental contexts by explicitly modeling cross-modal interactions. Our model builds on a multimodal diffusion transformer and fuses transcript-aligned speech latent with text-conditioned environmental context via joint attention. To enhance semantic consistency, we introduce a domain-specific representation alignment objective tailored to environment-aware TTS, leveraging complementary self-supervised representations from speech and audio encoders. Experimental results show that ImmersiveTTS achieves higher naturalness, intelligibility, and audio fidelity than existing approaches across objective metrics and human listening tests.

1 Introduction

Text-guided audio generation has emerged as a prominent research area in speech and audio processing, broadly categorized into sub-tasks: text-to-audio (TTA) and text-to-speech (TTS). TTA (Liu et al., 2023a, 2024; Kreuk et al., 2023; Huang et al., 2023) focuses on synthesizing non-speech audio, including foley effects, music, and environmental soundscapes, from natural language descriptions. Because most TTA models are not designed or optimized to capture fine-grained phonetic and prosodic structure, they often struggle to synthesize intelligible speech with precise linguistic content, even when given instructions such as "A woman is speaking."

In contrast, TTS (Ren et al., 2021; Kim et al., 2021a; Lee et al., 2022, 2025; Kim et al., 2025; Chen et al., 2025a) aims to generate natural-sounding human speech waveforms from textual input, such as characters, phonemes, or words. Despite its success, robust speech generation across diverse acoustic environments remains a significant challenge in TTS research. This difficulty stems from two main factors: (i) speech and environmental audio have largely been modeled in separate pipelines; and (ii) jointly generating heterogeneous audio sub-modalities, such as intelligible speech with environmental audio, within a single model is inherently complex, due to substantial differences in their acoustic structures and modeling requirements.

Motivated by these challenges, several recent studies have explored unified modeling for multiple audio generation tasks (Vyas et al., 2023; Yang et al., 2024; Liu et al., 2024; Choi et al., 2024). These approaches leverage powerful generative backbones such as diffusion models (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022), conditional flow matching (Lipman et al., 2023; Liu et al., 2023c; Yun et al., 2025), and language models (Chen et al., 2025a). Nevertheless, many of these systems optimize for each task separately and still struggle to synthesize natural speech with well-integrated environmental audio.

In particular, environment-aware TTS (Lee et al., 2024; Jung et al., 2025) has been explored as a method for generating speech and its surrounding acoustic context simultaneously. However, existing methods do not fully capture cross-modal interactions between speech and environmental audio. As a result, the synthesized outputs often exhibit speech-environment mismatch, leaving substantial room for improvement in overall coherence and immersion.

In this paper, we propose ImmersiveTTS, an environment-aware TTS model that jointly synthe-

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38295–38314

sizes natural speech with environmental audio by explicitly modeling interactions between linguistic content and environmental context, thus addressing these limitations. To enable this, we build on the multimodal diffusion transformer (MM-DiT) architecture (Esser et al., 2024), which was originally designed for image-text integration with a dual-stream backbone. We extend this approach to speech synthesis. Specifically, we assign transcript-aligned speech features and text-conditioned environmental context to parallel streams. We use joint attention between the two streams to explicitly model cross-modal interactions.

Although the MM-DiT architecture supports multimodal joint training, relying solely on its generative objective may be insufficient to learn speech representations that are both linguistically precise and grounded in environmental context. To stabilize cross-modal learning and improve semantic consistency, we introduce a domain-specific representation alignment objective tailored to environment-aware TTS, inspired by representation alignment (REPA) (Yu et al., 2025). Experimental results show that ImmersiveTTS achieves higher naturalness, intelligibility, and speech-environment coherence than existing methods. Ablation studies further validate the effectiveness of domain-specific alignment for environment-aware TTS. Audio samples and code implementations are provided at `https://jjunak-yun.github.io/ImmersiveTTS`.

2 Related Work

2.1 Environment-Aware Text-to-Speech

Environment-aware TTS aims to generate speech that matches a target acoustic environment, such as a background noise condition or a sound scene. Existing approaches can be organized according to how they obtain environmental information.

The first type of methods (Tan et al., 2022; Lu et al., 2025a; Glazer et al., 2025; Lu et al., 2025b) infer the acoustic context from reference audio and condition the TTS system on speaker and environmental attributes derived from it, either encoded as embeddings or used directly. In particular, (Tan et al., 2022) extends a Tacotron (Shen et al., 2018) by introducing separate encoders for speaker identity and room acoustics, while IDEA-TTS (Lu et al., 2025a), based on VITS (Kim et al., 2021a), incrementally disentangles speaker, content, and environment factors from reference speech. Building on a flow matching TTS backbone (Chen et al., 2025b), UmbraTTS (Glazer et al., 2025) introduces speech-to-environment ratio conditioning, while DAIEN-TTS (Lu et al., 2025b) uses a pre-trained speech-environment separation module for environment-aware TTS.

The second type of methods (Lee et al., 2024; Jung et al., 2025) take natural language prompts as input to describe the target acoustic scene, rather than relying on reference audio. Extending the AudioLDM framework, VoiceLDM (Lee et al., 2024) conditions a U-Net on two natural language prompts. A description prompt is encoded by a frozen CLAP encoder (Wu et al., 2023), while a content prompt is encoded by a SpeechT5 encoder (Ao et al., 2022). The resulting embeddings are injected into the U-Net via cross-attention. More recently, VoiceDiT (Jung et al., 2025) adopts Diffusion transformers (DiTs) with adaptive layer normalization (AdaLN) for environmental conditioning, enabling environment-aware speech synthesis from both text and visual prompts. Related studies, such as ViT-TTS (Liu et al., 2023b) and M2SE-VTTS (Liu et al., 2025b), also explore visual or spatial cues as additional modalities for TTS.

Compared with reference audio-based approaches, specifying the environment via text prompts offers advantages: it scales more naturally to arbitrary or unseen acoustic scenes and obviates the need to collect reference recordings. In this work, we focus on the latter strategy: utilizing text prompts to directly specify the desired environmental context. Despite these advantages, effectively fusing distinct textual cues, namely speech transcriptions and environmental descriptions, into a seamlessly integrated audio waveform remains a non-trivial modeling challenge.

2.2 Multimodal Diffusion Transformers

DiTs (Peebles and Xie, 2023) have been introduced as scalable alternatives to conventional U-Net backbones (Ronneberger et al., 2015) in diffusion models. The MM-DiT architecture proposed in SD3 (Esser et al., 2024) extends DiT to the multimodal setting by mapping text tokens and image patches into a unified token sequence and applying self-attention over all tokens. This approach allows for bidirectional cross-modal interaction at every layer within a single transformer. To accommodate modality-specific properties, MM-DiT adopts a dual-stream design that maintains separate representation paths for image and text tokens. In addi-

tion, multiple text encoders, such as CLIP (Radford et al., 2021) and T5 (Raffel et al., 2020), are supported within this unified design.

Following these advancements, recent audio generative models (Fei et al., 2024; Hung et al., 2024; Li et al., 2025; Cheng et al., 2025; Liu et al., 2025a; Wang et al., 2025; Shan et al., 2025) have adopted MM-DiT to condition audio generation on textual prompts and other contextual information in the latent space. These developments underscore the flexibility and effectiveness of MM-DiT as a backbone for multimodal audio generation. In this work, we adapt the MM-DiT architecture for environment-aware TTS. Unlike previous general audio models, we specialize its dual-stream design to treat transcript-aligned speech features and environmental cues as distinct yet interacting modalities, thereby facilitating precise linguistic control within immersive acoustic scenes.

2.3 Representation Alignment

The REPA method (Yu et al., 2025) regularizes diffusion transformers by aligning intermediate hidden states of DiT with the features produced by a powerful self-supervised learning (SSL) teacher encoder (Oquab et al., 2023). It is designed to improve semantic fidelity and accelerate convergence in diffusion and flow matching models.

Although REPA was first introduced in the context of image generation, subsequent work has begun to adopt REPA-based objectives for TTS and TTA tasks. ACE-Step (Gong et al., 2025) incorporates a semantic alignment loss, aligning intermediate features from its Linear DiT (Xie et al., 2025) with representations from pretrained MERT (Li et al., 2023) and mHuBERT (Boito et al., 2024). Vevo2 (Zhang et al., 2025) adopts REPA in its flow matching acoustic model, aligning an intermediate representation with W2v-BERT 2.0 features (Chung et al., 2021) to improve training efficiency and controllability of speech and singing voice generation. A-DMA (Choi et al., 2025) introduces text and speech-guided alignment losses using a CTC (Graves et al., 2006) and a speech SSL model such as HuBERT (Hsu et al., 2021), and shows that these alignment objectives accelerate convergence and improve speech quality. Building on these insights, our approach employs a domain-specific alignment scheme that uses separate pretrained SSL encoders to capture the distinct properties of speech and the environment. We elaborate on this architectural design in the following section.

3 ImmersiveTTS

In this section, we present ImmersiveTTS, an environment-aware TTS model built on the MM-DiT backbone to capture the interplay between speech and environmental context. For high-fidelity generation and stable training, we adopt a flow matching generative objective coupled with a domain-specific REPA. The overall pipeline is illustrated in Figure 1, and the details of each component are described below.

3.1 Preliminaries on Flow Matching

Flow matching or rectified flow (Lipman et al., 2023; Liu et al., 2023c) provides an approach that aims to learn a transformation between simple prior π_0 and data distribution π_1 on $\mathbb{R}^{d_z}$. The transformation is expressed as the following ordinary differential equation (ODE) over time $t \in [0, 1]$:

$$\frac{\mathrm{d}}{\mathrm{d}t} Z_t = v(Z_t, t), \quad Z_0 \sim \pi_0, \quad Z_1 \sim \pi_1, \quad (1)$$

where $v : \mathbb{R}^{d_z} \times [0, 1] \to \mathbb{R}^{d_z}$ is the time-dependent velocity field and π_0 typically follows a standard Gaussian distribution $\mathcal{N}(0, I)$.

We parameterize the field with a neural network v_θ. It is trained by minimizing the mean squared error between the velocity of straight paths connecting random pairs (Z_0, Z_1) and the neural velocity as follows:

$$\mathcal{L}_{\text{Flow}}(\theta) = \mathbb{E}_{t,Z_0,Z_1}\left[||(Z_1 - Z_0) - v_\theta(Z_t, t)||^2\right], \quad (2)$$

where $Z_t = (1 - t)Z_0 + tZ_1$ represents the linear interpolation between $Z_0 \sim \pi_0$ and $Z_1 \sim \pi_1$, and $t \in [0, 1]$ denotes the time step. Given the learned velocity field v_θ, the flow-based model transports samples from the prior π_0 to the target distribution π_1 along straight trajectories.

3.2 Audio Compression

To capture both speech and general audio characteristics within a unified latent space, we employ the pretrained variational autoencoder (VAE) used in AudioLDM2 (Liu et al., 2024). Let $X_{\text{wav}} \in \mathbb{R}^{d \cdot f_s}$ denote the raw waveform of duration d seconds with a sampling rate of f_s. We first convert X_{wav} into a log-mel spectrogram $X_{\text{mel}} \in \mathbb{R}^{F \times L}$, where F is the number of mel bins and L is the length of mel-spectrogram sequence. The VAE encoder compresses X_{mel} into a latent representation $Z \in \mathbb{R}^{8 \times F/4 \times L/4}$ by downsampling the time-frequency axes by a factor of 4. The VAE decoder

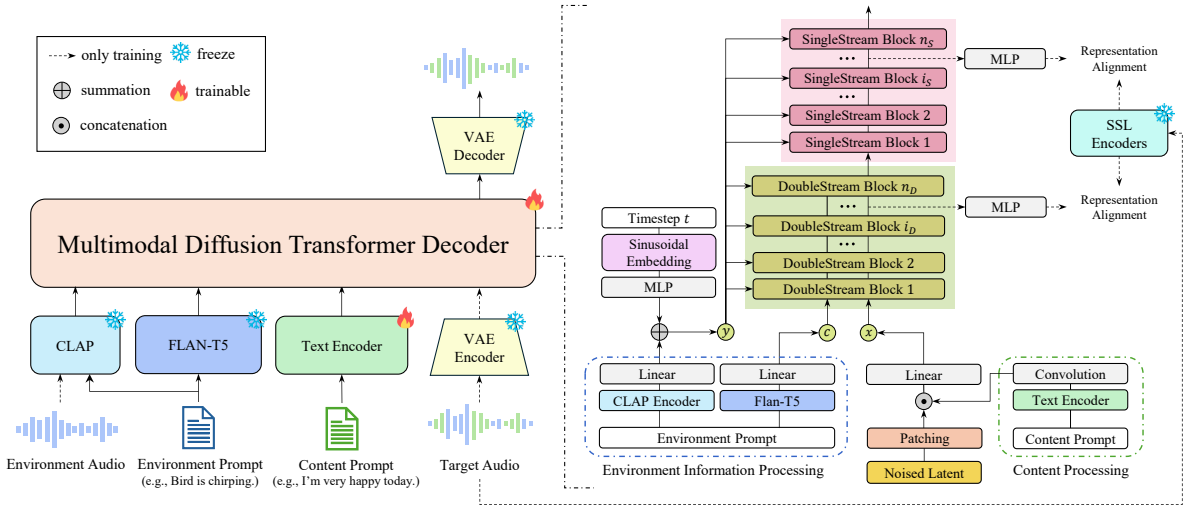

Figure 1: Overview of ImmersiveTTS. A dual-stream MM-DiT backbone conditions the speech stream on content prompt-aligned linguistic features. At the same time, Flan-T5 token embeddings drive the environmental context stream, and CLAP embeddings modulate AdaLN for global conditioning. The model is trained with flow matching and domain-specific REPA objectives.

reconstructs $\hat{X}_{\text{mel}}$ from Z, and a pretrained vocoder (Kong et al., 2020; Kim et al., 2021b) converts $\hat{X}_{\text{mel}}$ into the waveform $\hat{X}_{\text{wav}}$. All VAE parameters are frozen during training.

3.3 Multimodal Diffusion Transformer for Environment-Aware Text-to-Speech

Our objective is to generate speech that simultaneously preserves linguistic content and aligns with an environmental context. Accordingly, the model is conditioned on two distinct textual inputs: a content prompt y_{cont} (i.e., the transcription) and an environment prompt y_{env} (i.e., the background description). To effectively model the interplay between speech latents and environmental cues, we employ the MM-DiT backbone (Esser et al., 2024), which accommodates heterogeneous inputs and provides a robust foundation for cross-modal fusion.

In particular, we adopt the Flux architecture[1] to synthesize high-fidelity speech. The architecture comprises a stack of double-stream DiT layers followed by single-stream DiT layers. In the double-stream stage, we decouple the processing into two dedicated pathways: (i) the environmental context stream, which encodes the fine-grained environmental context tokens derived from y_{env}, and (ii) the speech stream, which processes the noisy audio latents Z_t conditioned on the linguistic features from y_{cont}. Crucially, these parallel streams exchange information through joint attention mechanisms. The detailed internal flow of the double-stream DiT block is illustrated in Appendix A. This design allows the speech generation process to dynamically attend to and harmonize with the environmental cues without losing its linguistic structure. Subsequently, only the representations from the speech stream are forwarded to the single-stream blocks, where they are further refined via self-attention layers. We describe the detailed configuration of each stream below.

Environmental Context Stream. For audio generation, existing approaches typically rely on either CLAP (Wu et al., 2023) encoders for coarse, global sound semantics or T5-family (Raffel et al., 2020) encoders for fine-grained detail. We adopt a dual-granularity conditioning strategy that leverages the complementary strengths of both the CLAP and T5 encoders (Xue et al., 2024).

First, to capture the global acoustic context, we project the CLAP embedding from y_{env} using an MLP and combine it with the diffusion timestep embedding to condition the AdaLN modules. By modulating the AdaLN scale and shift parameters (γ, β) across transformer blocks, it globally conditions the generation process.

In parallel, we apply a linear projection to token-level T5 embeddings of y_{env} to match the model dimension, and feed them into the environment context stream as an input sequence. This allows the speech stream to selectively attend to local environmental details through the joint attention mechanism in the double-stream layers. This ap-

[1] https://github.com/black-forest-labs/flux

proach balances global semantic consistency with fine-grained acoustic fidelity.

Environment-Aware Speech Stream. To synthesize intelligible speech that faithfully follows the content prompt y_{cont}, we incorporate an explicit temporal alignment that directly injects linguistic features into the speech stream. Following the framework of (Kim et al., 2020), the text encoder converts y_{cont} into a hidden representation $\tilde{\mu}_{1:L}$, while the monotonic alignment search (MAS) algorithm estimates phone-level durations $d'_{1:L}$. The hidden vectors are then expanded based on d' to produce a frame-level prior mel representation μ. The text encoder and duration predictor are mainly optimized using the prior loss $\mathcal{L}_{\text{Prior}}$ and MAS-based duration loss $\mathcal{L}_{\text{Dur}}$ as in (Kim et al., 2020).

To align the prior representation μ with the audio latent space, μ is processed through a convolution network, which bridges the structural gap between the mel-spectrogram space and the VAE latent manifold (Jung et al., 2025). The resulting features are concatenated with the noisy latent Z_t along the channel dimension and fed into the environment-aware speech stream. Within the MM-DiT layers, this speech stream actively exchanges information with the sequence of the environment context stream via joint attention. After passing through the full stack of double-stream blocks, only the environmentally-adapted speech representations are forwarded to the single-stream blocks for high-fidelity refinement.

3.4 Domain-Specific Representation Alignment

Without explicit feature-level alignment during training, we find that the diffusion backbone often struggles to simultaneously preserve linguistic intelligibility and environmental fidelity along the denoising trajectory, leading to content errors and acoustically inconsistent scenes. To enhance training stability and convergence, we extend the REPA strategy (Yu et al., 2025) to our multi-domain setting involving speech and environmental audio.

Domain-Specific SSL Encoders. For domain-specific REPA, we adopt a dual-teacher strategy that leverages the complementary strengths of specialized encoders. Building on the insights of (Chang et al., 2025), we use WavLM (Chen et al., 2022) and ATST-Frame (Li et al., 2024) as target encoders: (i) WavLM, a speech-specialized SSL model, selected to enforce precise phonetic and linguistic fidelity; and (ii) ATST-Frame, an audio-specialized SSL model, chosen to capture rich environmental acoustic events. Aligning to this heterogeneous pair rather than a single encoder encourages target representations that reflect both high-fidelity linguistic content and detailed environmental context.

Alignment Objective. Let $\{E_k\}_{k=1}^{K}$ denote a set of K pretrained SSL encoders. For a target audio input $X \sim p_{\text{data}}$, the kth encoder yields a target representation $r_k = E_k(X) \in \mathbb{R}^{B\times L_k\times D_k}$, where B denotes the batch size, and L_k, D_k represent the sequence length and dimensionality, respectively. To align our model with these targets, we extract hidden features $h_k \in \mathbb{R}^{B\times L_h\times D_h}$ specifically from the intermediate layers of the speech stream. These features are passed through a lightweight MLP projector to obtain $h'_k = \text{MLP}_k(h_k) \in \mathbb{R}^{B\times L_h\times D_k}$, mapping the transformer features into the encoder representation space. Following (Gong et al., 2025), we match the temporal resolutions of the projected features h'_k and target features r_k by interpolating or pooling them to a common temporal length $\tilde{L}$, yielding synchronized sequences $\tilde{h}'_k$ and $\tilde{r}_k$. The REPA loss is based on cosine similarity $\text{CosSim}(\cdot,\cdot)$ defined as follows:

$$\mathcal{L}_{\text{SSL}_k} = -\mathbb{E}_X\left[\text{CosSim}(\tilde{r}_k, \tilde{h}'_k)\right]. \quad (3)$$

Finally, the total objective is a weighted sum of the domain-specific alignment losses:

$$\mathcal{L}_{\text{REPA}} = \sum_{k=1}^{K} \lambda_k \mathcal{L}_{\text{SSL}_k}, \quad (4)$$

where λ_k is a hyperparameter to control the influence of each teacher. In our experiments, we set $\lambda_k = 1$ for all k.

3.5 Training and Inference

Training. The model is optimized with four losses during training. The velocity predictor and convolutional mapper are optimized with the flow matching objective $\mathcal{L}_{\text{Flow}}$ and the alignment objective $\mathcal{L}_{\text{REPA}}$. The text encoder and duration predictor receive gradients backpropagated through the conditioning pathway and, in addition, are directly supervised by the MAS-based prior loss $\mathcal{L}_{\text{Prior}}$ and the duration loss $\mathcal{L}_{\text{Dur}}$. Our final objective is

$$\mathcal{L} = \lambda_{\text{P}}\mathcal{L}_{\text{Prior}} + \lambda_{\text{D}}\mathcal{L}_{\text{Dur}} + \lambda_{\text{F}}\mathcal{L}_{\text{Flow}} + \lambda_{\text{R}}\mathcal{L}_{\text{REPA}}, \quad (5)$$

where we set all loss weights to 1 in our experiments. We freeze the CLAP and T5 encoders and draw the timestep $t \in (0, 1)$ from a logit-normal distribution with mean of 0 and variance of 1 (Esser et al., 2024), rather than uniformly from $\mathrm{U}(0, 1)$.

To enable flexible control over synthesized attributes, we adopt dual classifier-free guidance (CFG) (Ho and Salimans, 2022; Lee et al., 2024) by independently masking the content and environment prompt sequences with probability 0.1 during training.

Inference. During sampling, we first sample a random noise $Z_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The explicit velocity field is adjusted using dual CFG as follows:

$$\begin{aligned}\tilde{v}_\theta(Z_t, y_{\text{env}}, y_{\text{cont}}) &= v_\theta(Z_t, y_{\text{env}}, y_{\text{cont}}) \\ &+ \omega_{\text{env}}\Big(v_\theta(Z_t, y_{\text{env}}, \emptyset_{\text{cont}}) - v_\theta(Z_t, \emptyset_{\text{env}}, \emptyset_{\text{cont}})\Big) \\ &+ \omega_{\text{cont}}\Big(v_\theta(Z_t, \emptyset_{\text{env}}, y_{\text{cont}}) - v_\theta(Z_t, \emptyset_{\text{env}}, \emptyset_{\text{cont}})\Big),\end{aligned} \tag{6}$$

where ω_{env} and ω_{cont} denote the guidance scale of each modality, while $\emptyset_{\text{env}}$ and $\emptyset_{\text{cont}}$ denote the corresponding null conditions. We then employ Euler's method to solve the ODE in Equation 1:

$$Z_{t+\tau} = Z_t + \tau \cdot \tilde{v}_\theta(Z_t, t, y_{\text{env}}, y_{\text{cont}}). \tag{7}$$

Leveraging flow matching-based ODE sampling, we can generate high-quality latent features in fewer sampling steps. Finally, we decode the generated latents with the VAE decoder and synthesize waveforms using the pretrained vocoder.

4 Experiments

4.1 Experimental Setup

Datasets. To construct a robust training corpus for environment-aware TTS, we use two datasets: LibriTTS (Zen et al., 2019) for high-quality speech and WavCaps (Mei et al., 2024) for diverse environmental sounds. We use the *train-clean-360* subset of LibriTTS to provide clean linguistic content. WavCaps contains 400k audio clips; we explicitly filter out samples containing spoken content, retaining 340k non-speech clips to avoid overlapping speech. Following (Jung et al., 2025), we construct the training corpus by mixing clean speech from LibriTTS with environmental sounds from WavCaps. For each mixture, the environmental audio sample is mixed at a signal-to-noise ratio (SNR) value uniformly sampled between 2 and 10 dB. To ensure the model maintains the capability to generate clean speech, we skip this mixing process and use clean speech only with probability 0.15.

Preprocessing. All audio samples are downsampled to 16 kHz and converted into a mel-spectrogram with 64 mel bins using an STFT with an FFT size of 1024, window size of 1024, and hop length of 160. The frozen AudioLDM2[2] VAE then encodes this spectrogram into an 8-channel latent representation, which we use as the training target.

Implementation Details. ImmersiveTTS is trained for 400k steps on 2 NVIDIA RTX A6000 GPUs using the AdamW optimizer at a constant learning rate of 1×10^{-4}, and a batch size of 8 per GPU. The velocity field estimator consists of 12 double-stream blocks, 18 single-stream blocks, 6 attention heads, and a hidden state dimension of 1024, totaling approximately 450M trainable parameters. Detailed implementation information can be found in Appendix A.

4.2 Evaluation Metrics

We evaluate the performance of environment-aware TTS using both subjective and objective metrics. We conduct a mean opinion score (MOS) test to assess three aspects of the generated audio on a 5-point scale (1 to 5): speech naturalness (SN-MOS), environmental consistency (EC-MOS), and overall integration naturalness (ON-MOS). Detailed information on the MOS can be found in Appendix B.

For objective evaluation, we employ the word error rate (WER) to assess speech intelligibility and content accuracy. For WER, we use Whisper-*Large-v3* (Radford et al., 2023) as the automatic speech recognition (ASR) model. We additionally measure speaker similarity using speaker embedding cosine similarity (SECS) computed with WavLM-*base-sv* (Chen et al., 2022), which serves as an objective metric for speaker identity preservation. We report the Frechet audio distance (FAD) to measure the distribution distance between generated and target audio using VGGish (Hershey et al., 2017), and the CLAP score to quantify text-audio coherence with the environment description, defined as the cosine similarity between the CLAP embeddings of the text prompt and the synthesized audio. We also report the number of function evaluations (NFEs) as a measure of sampling efficiency.

[2] https://huggingface.co/cvssp/audioldm2

Model	#Param.	NFEs	Subjective			Objective		
			SN-MOS(↑)	EC-MOS(↑)	ON-MOS(↑)	WER(↓)	FAD(↓)	CLAP(↑)
Ground Truth	-	-	-	-	-	22.29	-	0.503
Reconstructed	-	-	4.08 ± 0.08	4.16 ± 0.08	3.49 ± 0.05	22.58	-	0.488
VoiceLDM (Lee et al., 2024)	508M	200	3.41 ± 0.06	3.33 ± 0.07	2.55 ± 0.05	16.45	8.75	0.229
VoiceDiT (Jung et al., 2025)	566M	200	3.47 ± 0.05	3.44 ± 0.07	2.63 ± 0.05	11.68	9.07	0.263
ImmersiveTTS	450M	25	**4.20** ± 0.07	**3.48** ± 0.07	**3.47** ± 0.05	**8.06**	**5.80**	**0.308**

Table 1: Experimental results for environment-aware text-to-speech on the AudioCaps *test* set. **#Param.** denotes the number of trainable parameters. The MOS results are reported with a 95% confidence interval.

Model	#Param.	NFEs	Subjective			Objective		
			SN-MOS(↑)	EC-MOS(↑)	ON-MOS(↑)	WER(↓)	FAD(↓)	CLAP(↑)
Ground Truth (Augmented)	-	-	-	-	-	7.86	-	0.317
Reconstructed	-	-	4.02 ± 0.08	3.95 ± 0.08	3.41 ± 0.07	3.59	-	0.291
VoiceLDM (Lee et al., 2024)	508M	200	3.32 ± 0.06	3.24 ± 0.07	2.91 ± 0.08	11.20	6.98	0.118
VoiceDiT (Jung et al., 2025)	566M	200	3.45 ± 0.06	**3.38** ± 0.06	3.12 ± 0.08	7.08	5.37	0.134
ImmersiveTTS	450M	25	**4.18** ± 0.07	3.32 ± 0.06	**3.23** ± 0.08	**4.48**	**3.92**	**0.207**

Table 2: Experimental results for environment-aware text-to-speech on the augmented test set with Seed-TTS *test-en* and AudioCaps *test* sets. The MOS results are reported with a 95% confidence interval.

5 Results

5.1 Main Results

We compare ImmersiveTTS with diffusion-based environment-aware TTS models, VoiceLDM (Lee et al., 2024) and VoiceDiT (Jung et al., 2025), which condition generation on natural language environment descriptions via a pretrained CLAP text encoder. For a fair comparison, all models are trained from scratch on the same training corpus with the same number of optimization steps. We also select optimal dual CFG scales based on preliminary tuning on an evaluation set.[3]

Table 1 reports the performance on the AudioCaps *test* set, where the ground truth audio contains both speech and background audio. *Reconstructed* denotes the reconstruction of the target audio obtained by encoding and decoding it through the STFT, VAE, and vocoder following Section 3.2. We use *Reconstructed* as the practical upper bound for subjective evaluation. We note that relatively high WER for ground-truth and reconstructed samples has also been reported in prior research (Lee et al., 2024; Jung et al., 2025; Lu et al., 2025b), as ASR can degrade when background audio partially masks speech.

Compared to existing approaches, ImmersiveTTS achieves substantially higher subjective scores, especially in SN-MOS and ON-MOS, indicating that our model generates more natural speech that is better integrated into the acoustic environment. ImmersiveTTS also improves objective metrics, yielding lower WER and FAD and higher CLAP score, suggesting better intelligibility, audio quality, and text-audio semantic alignment. It validates that our joint attention and domain-specific REPA strategy successfully preserves linguistic clarity while ensuring semantic alignment with the environment.

Model	NFEs	WER(↓)	UTMOS(↑)	SECS(↑)	FAD(↓)	CLAP(↑)
Ground Truth	-	2.21	4.00	0.9218	-	0.512
VoiceLDM	200	14.01	2.82	0.7601	8.71	0.288
VoiceDiT	200	11.08	**3.33**	**0.8942**	8.23	0.302
ImmersiveTTS	25	**9.89**	3.23	0.8859	**7.81**	**0.323**

Table 3: Objective evaluation results for single task on the LibriTTS *test* and AudioCaps *test* set.

Table 2 further evaluates models on an augmented test set constructed from Seed-TTS *test-en* and non-speech clips from AudioCaps *test* set, where clean speech is mixed with environmental audio. Although VoiceDiT obtains a slightly higher EC-MOS, ImmersiveTTS achieves the best SN-MOS and ON-MOS, indicating that it simultaneously achieves stronger overall naturalness and speech-environment integration. ImmersiveTTS also achieves the lowest WER and improves FAD and CLAP compared to existing methods. Notably, ImmersiveTTS attains these gains with only 25 NFEs, compared to 200 for the diffusion baselines. Overall, ImmersiveTTS shows consistent improvements in both real and augmented test sets. Additional evaluation of speaker similarity and broader baseline comparisons are provided in Appendix E and Appendix F, respectively.

[3]We use $(\omega_{\text{env}}, \omega_{\text{cont}}) = (3, 5)$ for VoiceLDM and $(5, 3)$ for VoiceDiT.

Strategy	Target SSL	Domain		Speech	Environment	
		Speech	Env.	WER(↓)	FAD(↓)	CLAP(↑)
Base	-	-	-	11.21	9.64	0.236
Single	WavLM	✓	-	10.97	8.02	0.231
	ATST	-	✓	13.77	8.78	0.271
	USAD	✓	✓	9.04	7.93	0.239
Dual	WavLM, USAD	✓	✓	8.95	7.33	0.248
	USAD, ATST	✓	✓	8.94	8.20	0.266
	WavLM, ATST	✓	✓	**8.06**	**5.80**	**0.308**

Table 4: Experimental results across different teacher alignment strategies.

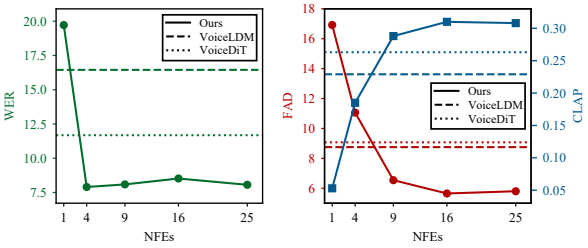

Figure 2: Comparison across different NFEs.

5.2 Single-Task Evaluation Results

In addition to our main evaluation, we provide single-task results on TTS and TTA. We additionally use UTMOS (Saeki et al., 2022) as an objective proxy for perceived speech naturalness. As shown in Table 3, ImmersiveTTS yields the lowest WER among the diffusion baselines, suggesting improved intelligibility under the same evaluation setting. In terms of speech naturalness, our system achieves UTMOS comparable to VoiceDiT while using substantially fewer sampling steps. The SECS results show that ImmersiveTTS preserves speaker identity better than VoiceLDM, although it remains slightly below VoiceDiT.

For TTA, ImmersiveTTS attains the best FAD among the baselines and achieves the highest CLAP score, which is closest to *Ground Truth*, indicating stronger text-audio semantic alignment. We attribute this gain to our MM-DiT structure, which incorporates a dual-granularity conditioning strategy and an audio domain alignment objective.

5.3 Analysis on Representation Alignment Strategy

To evaluate the effect of domain-specific REPA, we experiment with six teacher configurations: three single-teacher settings and three dual-teacher combinations. We use the two encoders adopted in our model, WavLM and ATST-Frame, and additionally include USAD (Chang et al., 2025), a unified speech-audio SSL encoder trained via distillation from these teachers. *Base* denotes our model trained without the REPA objective. We follow the same experimental setup as in Section 5.1 and report results on the AudioCaps *test* set.

We first examine the isolated effect of each teacher using a single-teacher strategy, as shown in Table 4. Using WavLM as the teacher yields a lower WER than the *Base* and ATST-Frame-only setting, demonstrating improved content fidelity driven by the speech-focused target. Although ATST-Frame degrades in intelligibility, it improves environment-related metrics, achieving the highest CLAP score and improving FAD over the *Base*, suggesting better alignment with environmental context. Because FAD is measured on the mixed waveform, stronger prompt alignment does not always translate into lower FAD, and we observe that WavLM tends to achieve lower FAD than ATST-Frame in this setting. USAD improves all three metrics over the *Base*, suggesting the benefit of guidance that covers both speech and environmental audio.

Based on these observations, we use a dual-teacher strategy and compare variants that include USAD with the domain-specific pair. As shown in Table 4, dual-teacher alignment mitigates the domain trade-off often observed in the single-teacher strategy. Notably, the pairing of WavLM and ATST-Frame achieves the best performance across all metrics, outperforming configurations that incorporate USAD. This indicates that the benefit of dual-teacher REPA comes not only from adding an additional target but also from choosing teachers with complementary strengths that provide more targeted guidance for speech content and environmental acoustics. Additional experiments are provided in Appendix C.

5.4 Analysis on Sampling Steps

We analyze how the number of sampling steps affects performance in environment-aware TTS. We follow the same experimental setting as in Section 5.1 with AudioCaps *test* set. Figure 2 shows consistent improvements as the sampling step increases. Increasing NFEs reduces both WER and FAD, while improving CLAP score, with the largest gains observed when moving from very few steps to moderate steps (e.g., $1 \rightarrow 9$ steps).

Notably, ImmersiveTTS matches or exceeds the diffusion baselines with far fewer sampling steps. With only 9 steps, it achieves lower WER and FAD, and higher CLAP scores than VoiceLDM and VoiceDiT, both of which use 200 NFEs. This highlights a favorable quality-efficiency trade-off,

delivering comparable or better quality with substantially fewer inference steps.

6 Conclusion

We presented ImmersiveTTS, an environment-aware text-to-speech framework that jointly synthesizes intelligible speech and environmental audio within a unified flow matching-based generative model. Built on an MM-DiT backbone, ImmersiveTTS explicitly models cross-modal interactions through a dual-stream stage that fuses transcript-aligned speech features with text-conditioned environmental cues via joint attention. To mitigate the domain mismatch between speech and environmental audio and to stabilize training, we further introduce a domain-specific REPA objective that aligns intermediate representations with distinct SSL teachers specialized for speech and environmental audio, respectively. Across evaluations on real and augmented environment-aware TTS benchmarks, ImmersiveTTS yields higher overall quality and semantic fidelity than existing approaches. Comprehensive analysis and ablation studies confirmed the effectiveness of the proposed dual-stream interaction design and domain-specific REPA for environment-aware TTS.

7 Limitations

Despite the effectiveness of ImmersiveTTS in jointly synthesizing speech with environmental audio, we acknowledge several limitations. First, our training relies primarily on synthetic mixtures of speech and environmental audio. While this enables scalable training, it may not fully capture the complex acoustic interactions present in large-scale recordings in the wild. We also note that robustness across varying SNR conditions and scene difficulty levels remains underexplored in the current work. While ImmersiveTTS ensures robust control over linguistic content and speaker identity through its dedicated modules, it currently lacks explicit control over paralinguistic attributes such as prosody, speaking style, or emotion. Incorporating these factors will be a crucial next step for producing expressive speech that fully aligns with both the target scene and the intended emotional expression. In future work, we aim to address these limitations by exploring large-scale real-world recordings and developing a method for the granular control of paralinguistic factors to achieve more immersive speech synthesis.

8 Acknowledgments

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) under the artificial intelligence graduate school program (Korea University) (No. RS-2019-II190079) and artificial intelligence star fellowship support program to nurture the best talents (IITP-2026-RS-2025-02304828) grant funded by the Korea government (MSIT).

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, and 1 others. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*, pages 5723–5738.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mhubert-147: A compact multilingual hubert model. In *Ann. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*.

Heng-Jui Chang, Saurabhchand Bhati, James Glass, and Alexander H Liu. 2025. Usad: Universal speech and audio representation via distillation. In *IEEE Autom. Speech Recognit. Underst. Workshop (ASRU)*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, pages 1505–1518.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2025a. Neural codec language models are zero-shot text to speech synthesizers. *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. 2025. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2024. Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion. In *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, volume 38, pages 17862–17870.

Jeongsoo Choi, Zhikang Niu, Ji-Hoon Kim, Chunhui Wang, Joon Son Chung, and Xie Chen. 2025. Accelerating diffusion-based text-to-speech model training with dual modality alignment. In *Ann. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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. In *IEEE Autom. Speech Recognit. Underst. Workshop (ASRU)*, pages 244–250.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

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*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis, 2024. In *Proc. Int. Conf. Mach. Learn. (ICML)*.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. 2024. Flux that plays music. *arXiv preprint arXiv:2409.00587*.

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

Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. 2025. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 369–376.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and 1 others. 2017. Cnn architectures for large-scale audio classification. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 33, pages 6840–6851.

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

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, 29:3451–3460.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *Proc. Int. Conf. Mach. Learn. (ICML)*.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. 2024. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*.

Jaemin Jung, Junseok Ahn, Chaeyoung Jung, Tan Dat Nguyen, Youngjoon Jang, and Joon Son Chung. 2025. Voicedit: Dual-condition diffusion transformer for environment-aware speech synthesis. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 33, pages 8067–8077.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021a. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 5530–5540.

Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, and Seong-Whan Lee. 2021b. Fre-gan: Adversarial frequency-consistent audio synthesis.

Seung-Bin Kim, Jun-Hyeok Cha, Hyung-Seok Oh, Heejin Choi, and Seong-Whan Lee. 2025. Fillerspeech: Towards human-like text-to-speech synthesis with filler insertion and filler style control. In *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, pages 34096–34113.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 17022–17033.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Textually guided audio generation. In *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Sang-Hoon Lee, Ha-Yeong Choi, Seung-Bin Kim, and Seong-Whan Lee. 2025. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *IEEE Transactions on Neural Networks and Learning Systems*.

Sang-Hoon Lee, Seung-Bin Kim, Ji-Hyun Lee, Eunwoo Song, Min-Jae Hwang, and Seong-Whan Lee. 2022. Hierspeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 16624–16636.

Yeonghyeon Lee, Inmo Yeon, Juhan Nam, and Joon Son Chung. 2024. Voiceldm: Text-to-speech with environmental context. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*.

Xian Li, Nian Shao, and Xiaofei Li. 2024. Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks. *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, pages 1336–1351.

Xiquan Li, Junxi Liu, Yuzhe Liang, Zhikang Niu, Wenxi Chen, and Xie Chen. 2025. Meanaudio: Fast and faithful text-to-audio generation with mean flows. *arXiv preprint arXiv:2508.06098*.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, and 1 others. 2023. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023a. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proc. Int. Conf. Mach. Learn. (ICML)*.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. 2024. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*, pages 2871–2883.

Huadai Liu, Rongjie Huang, Xuan Lin, Wenqiang Xu, Maozong Zheng, Hong Chen, Jinzheng He, and Zhou Zhao. 2023b. Vit-tts: visual text-to-speech with scalable diffusion transformer. In *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, pages 15957–15969.

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. 2025a. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*.

Rui Liu, Shuwei He, Yifan Hu, and Haizhou Li. 2025b. Multi-modal and multi-scale spatial environment understanding for immersive visual text-to-speech. In *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, volume 39, pages 24632–24640.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2023c. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Ye-Xin Lu, Hui-Peng Du, Zheng-Yan Sheng, Yang Ai, and Zhen-Hua Ling. 2025a. Incremental disentanglement for environment-aware zero-shot text-to-speech synthesis. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*.

Ye-Xin Lu, Yu Gu, Kun Wei, Hui-Peng Du, Yang Ai, and Zhen-Hua Ling. 2025b. Daien-tts: Disentangled audio infilling for environment-aware text-to-speech synthesis. *arXiv preprint arXiv:2509.14684*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, and Alaaeldin El-Nouby. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4195–4205.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn. (ICML)*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 28492–28518.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, pages 1–67.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (CVPR)*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. 2025. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, and 1 others. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*.

Yang Song and 1 others. 2021. Score-based generative modeling through stochastic differential equations. In *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Daxin Tan, Guangyan Zhang, and Tan Lee. 2022. Environment aware text-to-speech synthesis. In *Ann. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, and 1 others. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, and 1 others. 2025. Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and 1 others. 2025. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. 2024. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *IEEE/ACM Trans. Audio, Speech, Lang. Process. (TASLP)*.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, and 1 others. 2024. Uniaudio: An audio foundation model toward universal audio generation. In *Proc. Int. Conf. Mach. Learn. (ICML)*.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. 2025. Representation alignment for generation: Training diffusion transformers is easier than you think. In *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Jun-Hak Yun, Seung-Bin Kim, and Seong-Whan Lee. 2025. Flowhigh: Towards efficient and high-quality audio super-resolution with single-step flow matching. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu. 2025. Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning. *arXiv preprint arXiv:2508.16332*.

A Additional Implementation Details

To preserve speaker identity, we extract a speaker embedding from the speaker prompt using a pretrained WavLM-based speaker verification model[4]. This embedding is projected and fed as an additional conditioning input to the text encoder during both training and inference. For the environment prompt, we utilize the last hidden states of Flan-T5-*Large*[5] (Chung et al., 2024) as the token sequence for the environment context stream, while the global output vector of CLAP[6] provides coarse conditioning features. Figure 3 provides a zoomed-in view of the internal flow of the double-stream DiT block.

During inference, we set the classifier-free guidance scales to $\omega_{\text{env}} = 3$ and $\omega_{\text{cont}} = 3$ for each submodality. For the vocoder, we use the pretrained HiFi-GAN (Kong et al., 2020) to reconstruct the 16 kHz waveform from the sampled mel-spectrogram.

For representation alignment, we extract target features using WavLM-*Large*[7], ATST-Frame-*Base*[8], and USAD-*Base*[9]. We use the representations from the final layer of each encoder. Crucially, WavLM operates on clean speech from LibriTTS before mixing, ensuring that its alignment target focuses solely on linguistic fidelity. In contrast, ATST-Frame operates on mixed speech with environmental audio, allowing the target to capture the full acoustic scene. All teacher encoders remain frozen during training.

B Details of Subjective Evaluations

For the subjective evaluation, we conducted a mean opinion score (MOS) test to assess four aspects of the generated audio on a 5-point scale (1 to 5): speech naturalness (SN-MOS), environmental consistency (EC-MOS), overall integration naturalness (ON-MOS), and speaker similarity (S-MOS). SN-MOS measures the perceived naturalness of the synthesized speech; EC-MOS assesses how well the background audio matches the given environment description; ON-MOS evaluates overall naturalness, i.e., how naturally the speech and background audio are blended; and S-MOS measures how similar the synthesized speech is to the reference speaker in terms of speaker identity. To construct the evaluation data, we randomly sampled 30 utterances from each test set, excluding samples whose ground-truth audio contains multiple speakers. We didn't include the ground truth samples for the subjective evaluation, as they are perceptually very similar to the reconstructed samples produced by VAE and vocoder, which could lead to redundant comparisons and potentially bias listeners. All MOS ratings are reported with 95% confidence intervals.

We conducted these evaluations via crowdsourcing on Amazon Mechanical Turk[10]. Each evaluation was completed by 20 native English speakers residing in the United States. We allocated 42 USD

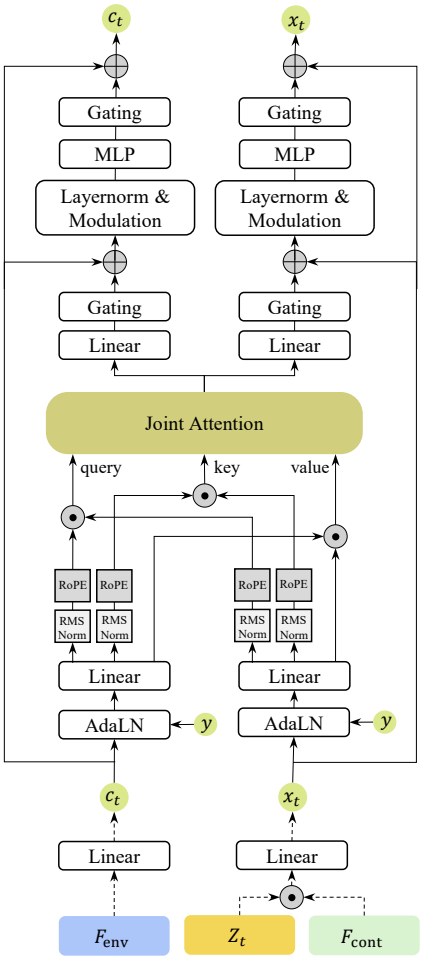

Figure 3: Illustration of the double-stream DiT blocks. Dashed lines indicate that this input initialization is applied only to the first double-stream block, where F_{env} denotes the Flan-T5 feature for the environment prompt, Z_t denotes the noisy latent at timestep t, and F_{cont} denotes the content-conditioned feature derived from the content encoder.

[4] https://huggingface.co/microsoft/wavlm-base-sv

[5] https://huggingface.co/google/flan-t5-large

[6] https://huggingface.co/laion/clap-htsat-unfused

[7] https://huggingface.co/microsoft/wavlm-large

[8] https://github.com/Audio-WestlakeU/audiossl

[9] https://huggingface.co/MIT-SLS/USAD-Base

[10] https://www.mturk.com/

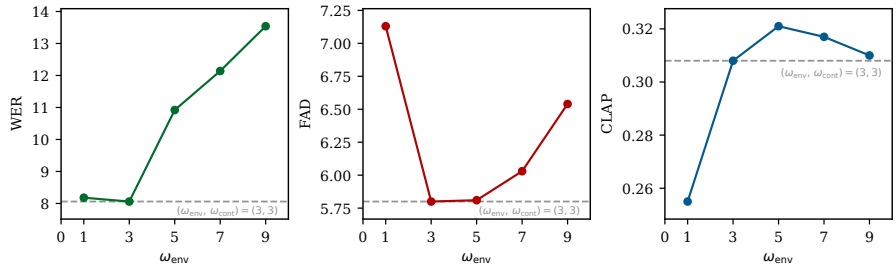

Figure 4: Objective evaluation results on AudioCaps *test* set across various environment guidance scales.

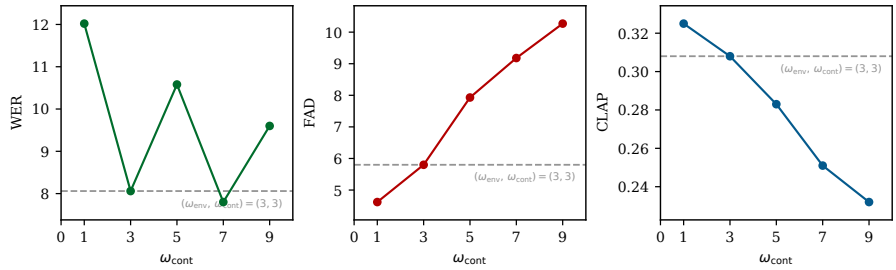

Figure 5: Objective evaluation results on AudioCaps *test* set across various content guidance scales.

Alignment Position	Target SSL	WER(↓)	FAD(↓)	CLAP(↑)
Single Teacher				
MM-DiT	WavLM	10.97	8.02	0.231
	ATST	13.77	8.78	0.271
	USAD	9.04	7.93	0.239
Single DiT	WavLM	9.94	7.83	0.226
	ATST	14.57	7.24	0.261
	USAD	10.47	7.02	0.266
Dual-Teacher				
MM-DiT & Single DiT	WavLM + USAD	9.31	7.76	0.231
	USAD + ATST	12.42	8.66	0.247
	WavLM + ATST	9.67	5.97	0.287
MM-DiT & MM-DiT	WavLM + USAD	8.95	7.33	0.248
	USAD + ATST	8.94	8.20	0.266
	WavLM + ATST	**8.06**	**5.80**	**0.308**

Table 5: Experimental results across different teachers and alignment positions on the AudioCaps *test* set.

per MOS task and ran the SN-MOS, EC-MOS, and ON-MOS assessments separately for each test set. In addition, the S-MOS evaluation on one test set costs 84 USD, resulting in a total cost of 336 USD. We additionally interspersed fake samples as attention checks throughout each evaluation set. We excluded ratings from participants who failed these checks. Screenshots of the Amazon Mechanical Turk interface for SN-MOS, EC-MOS, ON-MOS, S-MOS are shown in Figure 6, 7, 8, and 9.

C Analysis on Representation Alignment Injection Position

In this section, we report the effect of REPA injection position. As shown in Section 5.3, ImmersiveTTS establishes the benefit of complementary teachers. We check whether the same trend holds when the alignment loss is applied at different positions within the DiT backbone.

In our preliminary experiments, aligning WavLM at ten locations across the 30 DiT blocks did not reveal a clear monotonic pattern. However, aligning in the middle (or slightly earlier) layers was consistently more stable, consistent with prior findings (Yu et al., 2025). Based on this observation, we focus on the specific injection points: the 6th or 10th block in the MM-DiT stage (12 blocks) and the 4th block in the single DiT stage (18 blocks).

Table 5 reports objective results on the AudioCaps *test* set under different injection positions and teacher configurations. Overall, the qualitative trends observed in Section 5.3 remain consistent across the injection stage. In the single teacher setting, WavLM primarily improves speech accuracy, whereas ATST-Frame improves environment-related metrics, indicating that each teacher provides domain-specific semantic guidance. USAD exhibits a more balanced behavior depending on the injection stage. In the dual-teacher setting, combining WavLM and ATST-Frame consistently yields the best overall performance, and injecting both alignments within the MM-DiT blocks achieves the strongest results, suggesting robustness to this design choice.

Model	Reconstructed	VoiceLDM	VoiceDiT	ImmersiveTTS
S-MOS(↑)	3.18 ± 0.05	2.97 ± 0.06	**3.15** ± 0.06	**3.15** ± 0.06

Table 6: S-MOS results for environment-aware TTS on the Seed-TTS *test-en* and AudioCaps *test* sets.

Model	NFEs	WER(↓)	SECS(↑)	FAD(↓)	CLAP(↑)
Ground Truth	-	2.21	0.9218	-	0.512
VoiceLDM	200	14.01	0.7601	8.71	0.288
VoiceDiT	200	11.08	0.8942	8.23	0.302
CosyVoice2	10	**5.23**	**0.9227**	-	-
CosyVoice3	10	5.98	0.9152	-	-
AudioLDM2 (Audio)	200	-	-	3.79	0.407
TangoFlux	25	-	-	**2.96**	**0.502**
ImmersiveTTS	25	9.89	0.8859	7.81	0.323

Table 7: Objective evaluation results for single task on the LibriTTS *test* and AudioCaps *test* set with single task baselines.

D Analysis on Dual Classifier-Free Guidance Scale

To independently control sub-modality attributes, ImmersiveTTS adopts dual classifier-free guidance (CFG) with separate guidance scales for the environmental condition and the content condition. In the main experiments (Section 5.1), we use $(\omega_{\text{env}}, \omega_{\text{cont}}) = (3, 3)$. Here, to analyze the sensitivity to each guidance scale, we fix one scale to 3 and vary the other in $\{1, 3, 5, 7, 9\}$, reporting WER, FAD, and CLAP on the same evaluation setting.

Figure 4 varies ω_{env} with $\omega_{\text{cont}} = 3$. We observe that increasing ω_{env} beyond 3 substantially degrades intelligibility. WER increases from 8.06 at $\omega_{\text{env}} = 3$ to over 10.92 for $\omega_{\text{env}} \geq 5$. While FAD exhibits only mild variations, and CLAP shows a small improvement at moderate ω_{env} before plateauing at higher values. This suggests that overly strong environmental guidance can interfere with linguistic realization, even if it slightly improves text-audio alignment for the background. Conversely, setting $\omega_{\text{env}} = 1$ reduces semantic alignment and yields worse perceptual quality than $\omega_{\text{env}} = 3$.

Figure 5 varies ω_{cont} while fixing $\omega_{\text{env}} = 3$. The lowest WER is achieved at $\omega_{\text{cont}} = 7$. In contrast, as ω_{cont} increases, acoustic realism degrades steadily, reflected by a monotonic increase in FAD from 4.62 to 10.27. At the same time, semantic alignment decreases consistently, as CLAP drops from 0.325 to 0.232 over the same range. Overall, these results suggest that overly large ω_{cont} overemphasizes speech content, improving intelligibility only up to a point while harming overall audio quality and scene coherence.

Model	WER(↓)	FAD(↓)	CLAP(↑)
VoiceLDM	16.45	8.75	0.229
VoiceDiT	11.68	9.07	0.263
AudioLDM2 (Speech)	35.06	29.59	0.048
AudioLDM2 (Speech) + AudioLDM2 (Audio)	41.33	5.36	0.365
CosyVoice2 + TangoFlux	**6.76**	**4.01**	**0.452**
ImmersiveTTS	8.06	5.80	0.308

Table 8: Objective results for environment-aware TTS on the AudioCaps *test* set. The symbol '+' denotes mixing of separately generated speech and background audio.

Model	WER(↓)	FAD(↓)	CLAP(↑)
VoiceLDM	11.20	6.98	0.118
VoiceDiT	7.08	5.37	0.134
AudioLDM2 (Speech)	27.23	28.14	-0.069
AudioLDM2 (Speech) + AudioLDM2 (Audio)	35.78	3.39	0.236
CosyVoice2 + TangoFlux	**2.28**	**3.11**	**0.287**
ImmersiveTTS	4.48	3.92	0.207

Table 9: Objective results for environment-aware TTS on the Seed-TTS *test-en* and AudioCaps *test* sets. The symbol '+' denotes mixing of separately generated speech and background audio.

Overall, the analysis reveals a trade-off between speech clarity and semantic quality when scaling dual CFG. The balanced setting $(\omega_{\text{env}}, \omega_{\text{cont}}) = (3, 3)$ used in our main experiments provides a stable operating point, avoiding the sharp WER degradation observed with large ω_{env}, and the FAD and CLAP collapse observed with large ω_{cont}.

E Speaker Identity Preservation Evaluation Results

To further assess speaker identity preservation in the main environment-aware TTS setting, we conduct a S-MOS evaluation on the same augmented test set used in Table 2, where clean speech from Seed-TTS *test-en* is mixed with non-speech environmental audio from AudioCaps *test*. For each sample, we use the corresponding clean target speech waveform from Seed-TTS *test-en* as the reference speech for speaker similarity evaluation. As shown in Table 6, ImmersiveTTS achieves an S-MOS of 3.15, outperforming VoiceLDM and matching VoiceDiT. This score is also close to that of the reconstructed samples (3.18), suggesting that speaker identity is largely preserved in the environment-aware TTS setting.

F Broader Baseline Comparisons

F.1 Single-Task Baselines

To provide a broader comparison beyond the environment-aware TTS baselines in the main paper, we compare against CosyVoice2 (Du et al.,

2024) and CosyVoice3 (Du et al., 2025) for TTS, and against AudioLDM2-*Audio* (Liu et al., 2024) and TangoFlux (Hung et al., 2024) for TTA under the same evaluation settings as Table 3. Table 7 summarizes the results. As expected, single-task models perform strongly on their respective metrics. For TTS, CosyVoice2 and CosyVoice3 achieve substantially lower WER and higher SECS than the environment-aware TTS models, while for TTA, AudioLDM2 and TangoFlux achieve markedly better FAD and CLAP. These results reflect the inherent trade-off in environment-aware TTS, as a unified model must balance speech fidelity and background generation within a single system.

F.2 Mixing-based Pipeline Baselines

Building on the single-task baselines above, we further construct mixing-based pipeline baselines by separately generating speech and background audio and then mixing the two outputs. We consider two mixing-based baselines: one separately generates speech and audio using domain-specific AudioLDM2 checkpoints, and the other combines CosyVoice2 for speech with TangoFlux for background audio. Table 8 follows the same setting as Table 1, while Table 9 follows the augmented setting used in Table 2. As shown in both Table 8 and Table 9, AudioLDM2-based pipelines show relatively poor WER, mainly because the speech generated by AudioLDM2-*Speech* is less intelligible. The combination with CosyVoice2 and TangoFlux consistently performs the strongest across all objective metrics in both settings.

In contrast, such pipelines require separate generation and mixing, whereas ImmersiveTTS directly models speech-background interaction within a unified framework. We believe this unified formulation is important for improving coherence and realism by modeling how speech and background influence each other during synthesis.

G Potential Risks

The proposed environment-aware TTS system is designed to synthesize speech together with environmental audio based on provided textual prompt. As with other speech generation models, this capability inherently entails several potential risks. The system could be misused to generate unauthorized voice synthesis or deceptive audio content, potentially causing negative societal impact.

To mitigate these potential risks, our work is intended solely for research purposes, and we emphasized the importance of transparent disclosure of synthesized content and responsible use. Furthermore, when releasing our resources, we explicitly encourage users to adhere to these principles.

H AI Assistance

We used ChatGPT 5.2 to assist with proofreading and improving English grammar and expressions.

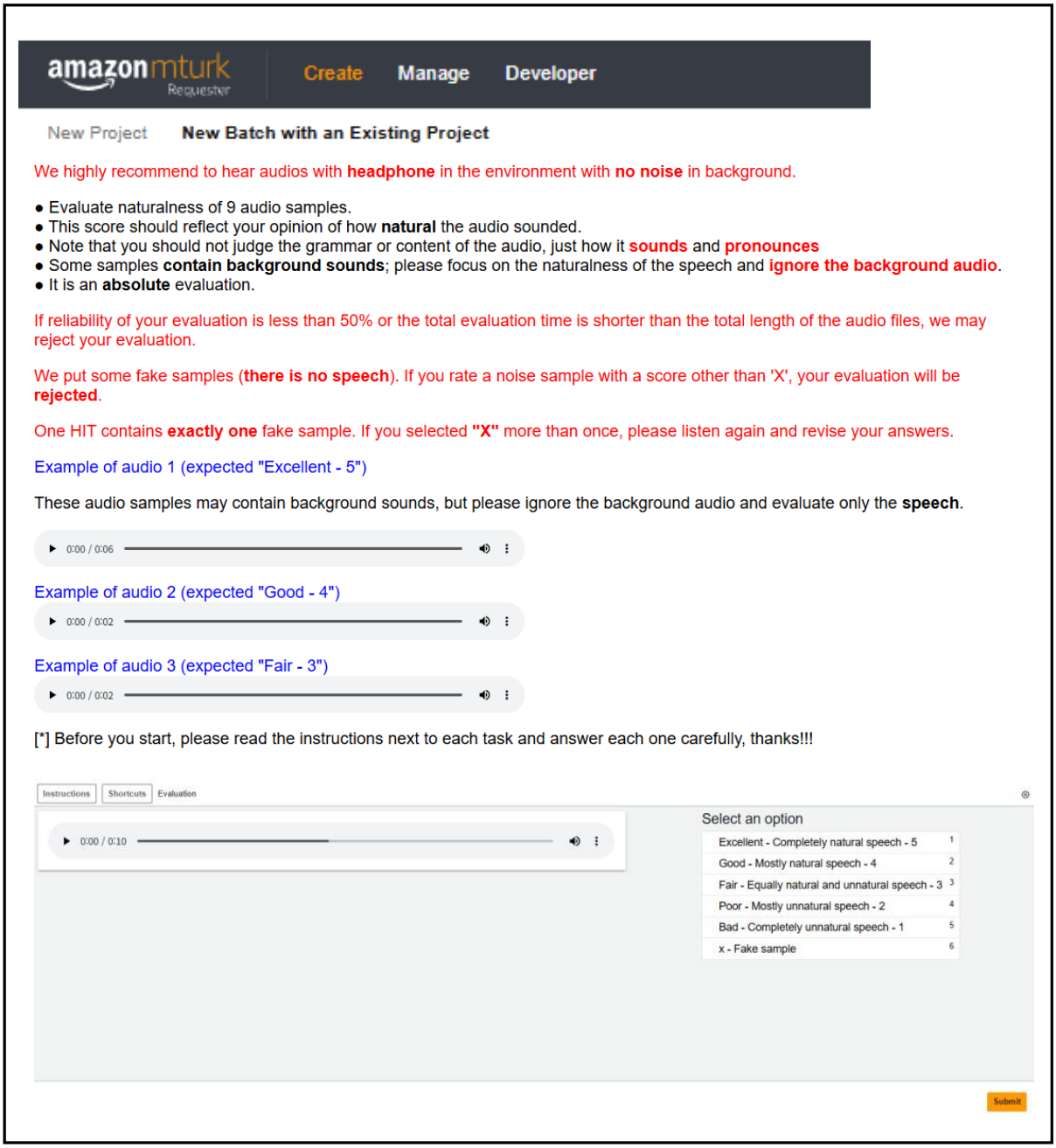

Figure 6: Detailed information on listener requirements and the SN-MOS evaluation interfaces.

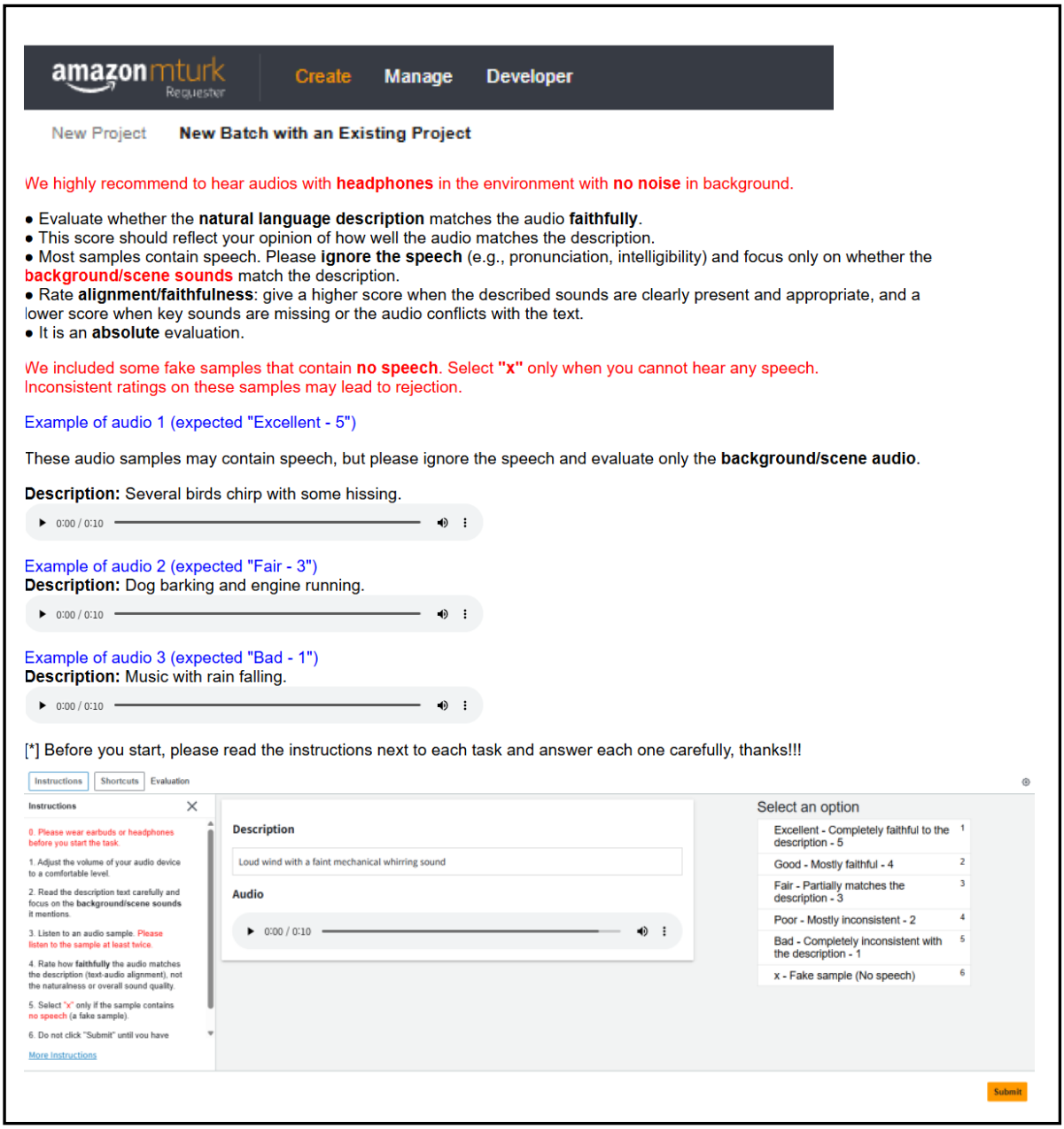

Figure 7: Detailed information on listener requirements and EC-MOS evaluation interfaces.

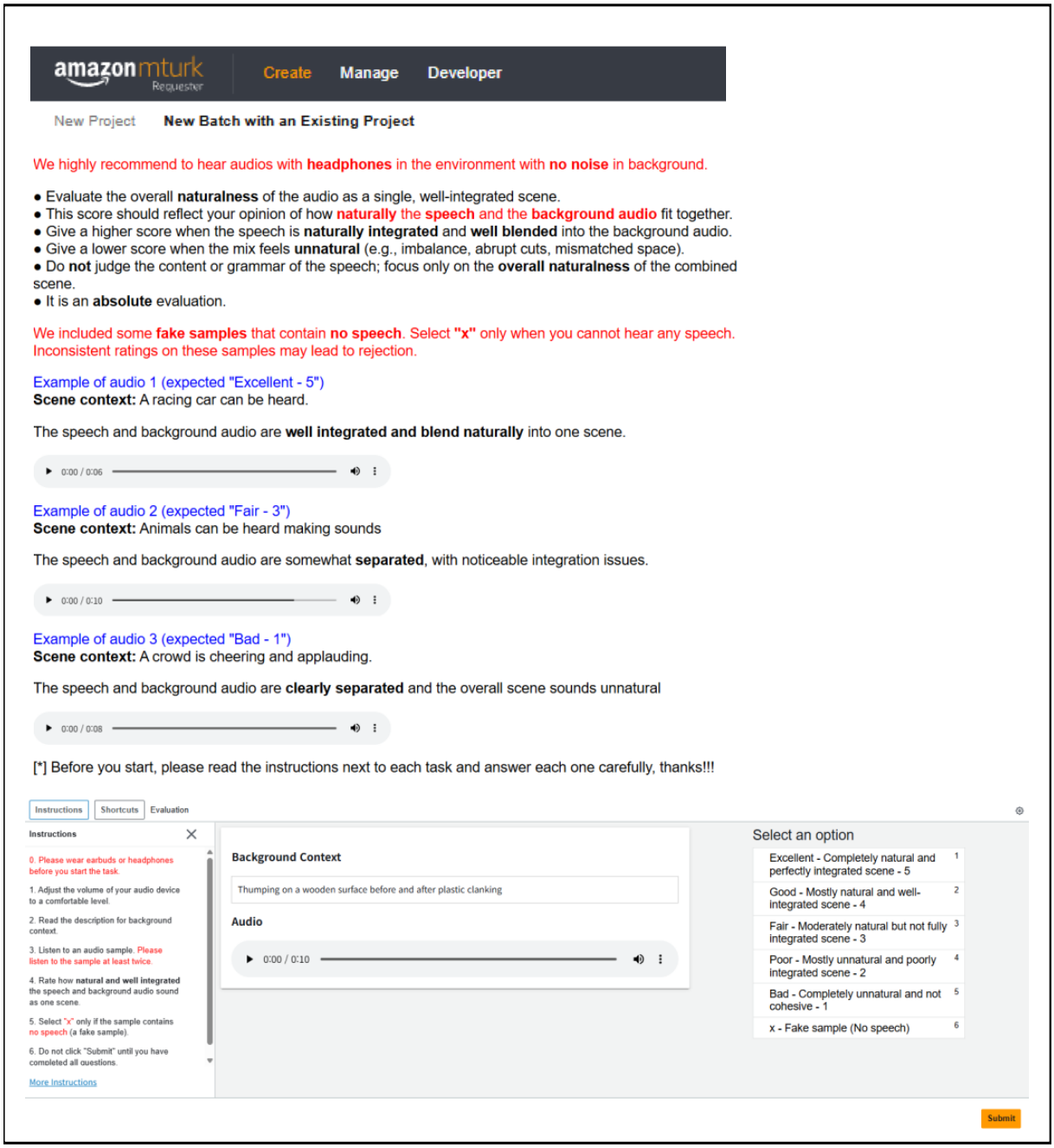

Figure 8: Detailed information on listener requirements and ON-MOS evaluation interfaces.

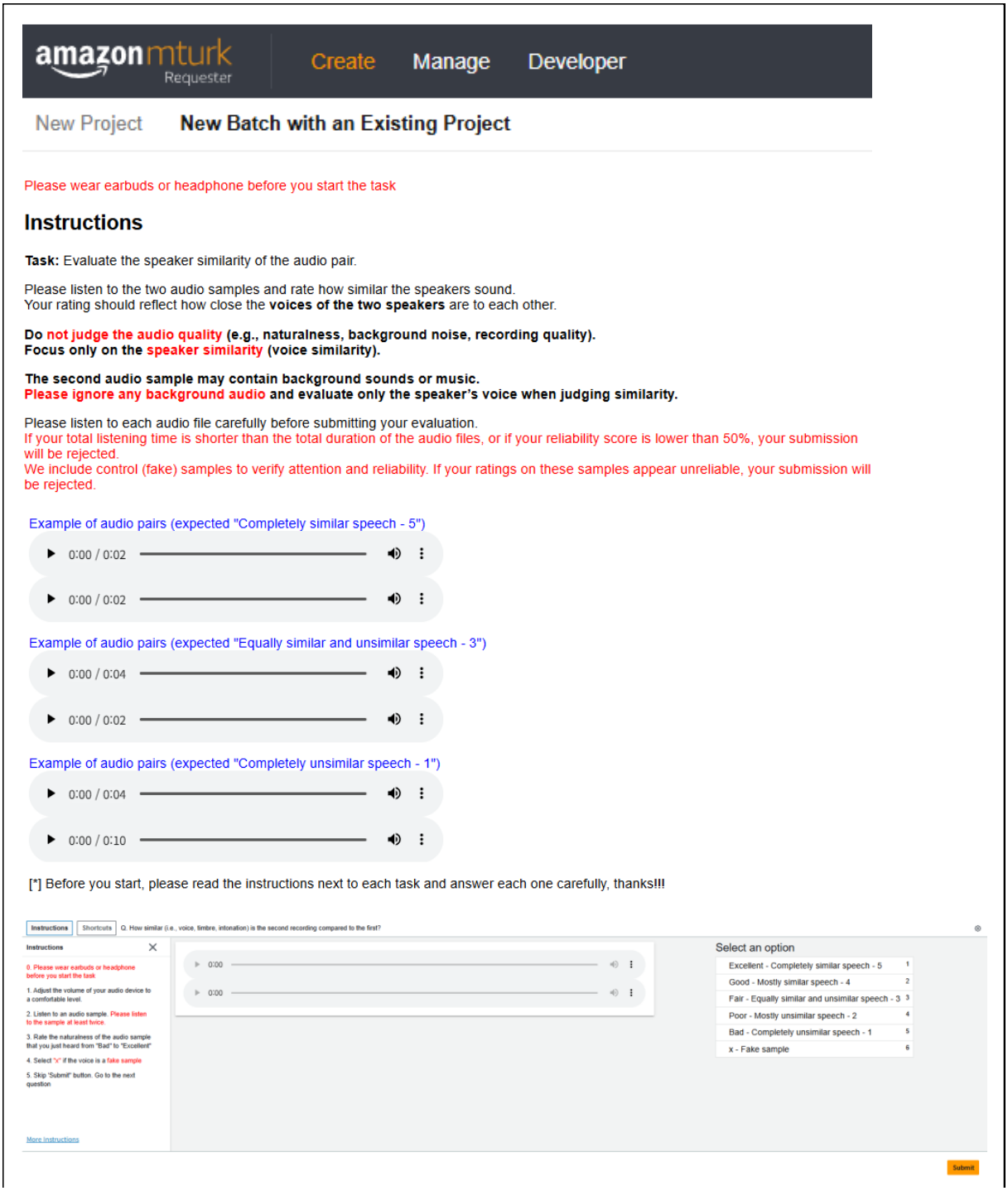

Figure 9: Detailed information on listener requirements and S-MOS evaluation interfaces.