Interspeech 2024
1-5 September 2024, Kos, Greece

Speak in the Scene: Diffusion-based Acoustic Scene Transfer toward Immersive Speech Generation

Miseul Kim[1,2], *Soo-Whan Chung*[1], *Youna Ji*[1], *Hong-Goo Kang*[2], *Min-Seok Choi*[1]

[1]Naver Cloud, South Korea, [2]Yonsei University, South Korea

miseul4345@dsp.yonsei.ac.kr, soowhan.chung@navercorp.com

Abstract

This paper introduces a novel task in generative speech processing, Acoustic Scene Transfer (AST), which aims to transfer acoustic scenes of speech signals to diverse environments. AST promises an immersive experience in speech perception by adapting the acoustic scene behind speech signals to desired environments. We propose AST-LDM for the AST task, which generates speech signals accompanied by the target acoustic scene of the reference prompt. Specifically, AST-LDM is a latent diffusion model conditioned by CLAP embeddings that describe target acoustic scenes in either audio or text modalities. The contributions of this paper include introducing the AST task and implementing its baseline model. For AST-LDM, we emphasize its core framework, which is to preserve the input speech and generate audio consistently with both the given speech and the target acoustic environment. Experiments, including objective and subjective tests, validate the feasibility and efficacy of our approach.

Index Terms: audio/speech generation, acoustic scene transfer, latent diffusion

1. Introduction

In our daily lives, we encounter a plethora of environmental sounds or acoustic scenes, including background noise, reverberation, and other auditory stimuli such as chirping birds, laughing children, or the murmur of distant conversations. The acoustic environment plays an essential role in human interactions, infusing conversations with vitality and often serving as a contextual cue for speech. For example, we often include supplementary sounds when conversing with others or reading a book to improve comprehension of the spoken content. Therefore, people often edit recorded speech with various sound samples, creating an immersive experience akin to speech recorded in a real environment.

Remarkable advances in the field of generative models enable the generation of high-quality speech and audio. Early generative methods, such as variational autoencoders (VAE) [1, 2] and generative adversarial networks (GAN) [3, 4], have successfully produced natural-sounding speech signals, surpassing the capabilities of previous discriminative methods. However, they often struggle with instability during training, which can lead to unpredictable and unsatisfactory outputs. Recent diffusion-based models have effectively addressed these challenges, markedly raising generation performance and achieving remarkable success in the extensive domains of audio and speech. Some studies [5, 6, 7] have underscored its efficacy in modeling the distribution of speech signals, especially in facilitating multi-speaker speech synthesis as well as high-quality synthesis. In the realm of music generation, diffusion has demonstrated its proficiency in crafting musical notes across various instruments such as piano, guitar, beats, and even vocals [8, 9, 10]. Furthermore, diffusion models can even generate general sounds by learning auditory characteristics from text description prompts and producing sound effects and foley sounds [11, 12, 13, 14]. Notably, AudioLDM [15, 16] demonstrates exceptional performance in audio generation, showcasing its ability to generate audio signals based on CLAP model [17, 18], retrieving acoustic characteristics from reference audio or text captions. VoiceLDM [19], a text-to-speech model, is built upon the framework of AudioLDM, and it generates speech signals with prompts and transcriptions to control speech, voice and various speaking environments. Similarly, DiffRENT [20] aims to transfer acoustic characteristics such as ambient noise and reverberation from reference speech to input speech signals.

In this paper, motivated by recent developments, we propose a novel speech processing task, **Acoustic Scene Transfer (AST)**, aimed at transferring the target acoustic scene onto an input speech signal. Although the framework of AST is similar to the foundational knowledge presented in DiffRENT, we further generalize the paradigm to encompass general audio rather than focusing solely on ambient noise. To the best of our knowledge, this is the first time such a task has been proposed for immersive speech generation, and we expect that its potential benefits across various speech processing fields. While this task may appear similar to adding background sound or room impulse responses as speech augmentation, AST does not simply replicate the acoustic scene of reference signals. Instead, it re-generates acoustic scenes, allowing for flexibility and diversity on generated scenes and relieving limitations imposed by noise or room impulse response in datasets. Therefore, AST would encourage more impactful and immersive speech generation beyond conventional natural voice generation.

We implement a latent diffusion model (LDM) for AST, **Acoustic Scene Transfer Latent Diffusion Model (AST-LDM)**, which predicts latent embeddings of a pre-trained variational autoencoder, following AudioLDM and VoiceLDM. Our proposed model mimics the background sound and acoustic characteristics of the target environment while retaining the spoken term and speaker identity present in the input speech. AST-LDM not only transfers acoustic backgrounds but also facilitates transitions from anechoic to noisy or reverberant environments, and vice versa. Furthermore, we demonstrate that AST-LDM can be conditioned not only by reference audios, as in DiffRENT, but also by text descriptions of acoustic scenes with a help of CLAP embeddings. By leveraging text prompts, the proposed model gains larger diversity and enriched representations of acoustic scenes, enabling it to generate unseen acoustic situations.

10.21437/Interspeech.2024-425

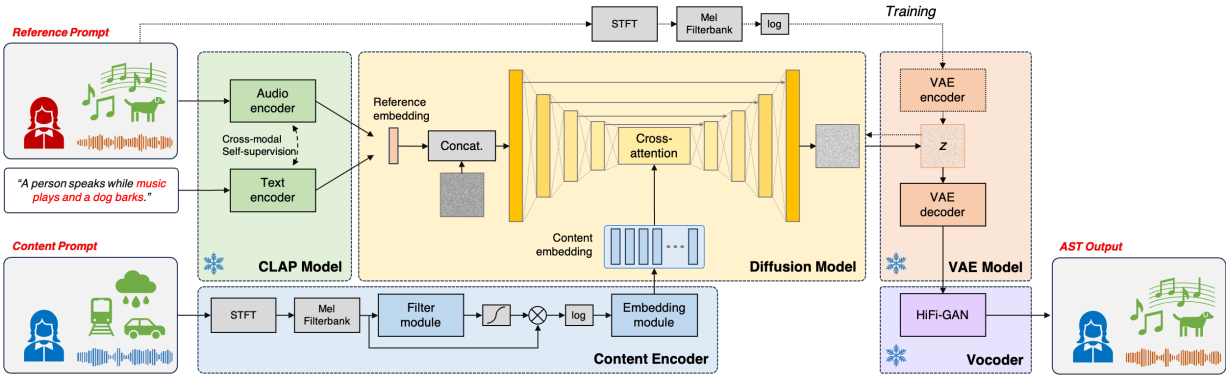


Figure 1: *Overview of the proposed model, AST-LDM. Dotted lines represent the training phase, while solid lines denote the inference process. Blocks marked with ❄ indicate pre-trained components that are not updated during training.*

2. Diffusion-based Audio Generation

Diffusion [21, 22] leads the field of generative learning by predicting the data distribution from Gaussian noise, and it allows the model to generate outputs with a large variability. The advent of latent diffusion [23], which operates in a latent space, has reduced excessive computational costs for processing raw data and increased the applicability of diffusion framework on various domains. In audio generation, diffusion models have achieved significant breakthroughs, as demonstrated in their capability in vision modality. Its application in speech synthesis led to the high-quality generation of speech waveforms conditioned by mel-spectrograms [24]. Beyond speech synthesis, diffusion techniques have been applied in speech enhancement [25, 26] and separation [27, 28], enabling the generation of clear speech from noisy inputs. Moreover, diffusion has demonstrated its versatility in creating a broad spectrum of audio content including natural sounds and instrumental music.

The method most closely aligned with ours is AudioLDM, which generates audio sounds by utilizing CLAP model to extract embeddings for the conditional generation. It adopts a self-supervised training approach; the model is trained to predict pre-trained latent embeddings for a broad spectrum of audio sounds, with the CLAP embedding of the audio itself serving as a reference prompt. During inference, the model employs the text encoder of CLAP model, using text descriptions of sound to steer the generation process rather than relying on the audio prompt. It is enabled by the common latent space between the audio and text encoders of CLAP model, which benefits from its training in cross-modal self-supervision. This approach is particularly effective in generating a varied range of audio outputs, including the ability to generate unseen audio outputs that are not present during the training stage. Inspired by AudioLDM, VoiceLDM [19] focuses on text-to-speech (TTS) synthesis, incorporating both text transcriptions and audio signals as conditions for training. It not only focuses on synthesizing clean speech but also simulates speech within the described environment. Leveraging CLAP model, VoiceLDM adjusts the acoustic scene based on textual scene descriptions, employing a dual classifier-free guidance method [29] to balance the influence of both conditions. DiffRENT [20] presents another innovative method, focusing on the transfer of acoustic scenes from reference to target speech. Based on a standard diffusion model, it directly estimates the target mel-spectrogram, employing an environmental encoder instead of CLAP model. The environmental encoder is jointly trained with the diffusion model, sharing the same training criteria. DiffRENT changes the ambient noise and room acoustics of the target speech to match those of the reference while preserving the spoken content.

In this work, we draw upon the training and conditioning strategies as well as frameworks of these models to generate optimal latent embeddings for AST. Our work aims to extend the boundaries of audio generation, leveraging the capabilities of diffusion models to create immersive speech experiences.

3. AST-LDM

In this section, we introduce our proposed model, AST-LDM, and outline its training strategy for the acoustic scene transfer task. The overall structure of AST-LDM is given in Figure 1. AST-LDM is conditioned on two prompts: the reference prompt, which consists of the target environments, and the content prompt, representing the input speech signal aimed at preserving the speaking contents in the output signal. Our model is inspired by the structure of AudioLDM which includes the diffusion model, CLAP encoder, VAE, and HiFi-GAN [4]. The VAE is charged with extracting the target latent embedding from the mel-spectrogram, while HiFi-GAN serves as the vocoder, converting the mel-spectrogram into an audio waveform. In the following subsections, we will offer a detailed demonstration of the modules of our proposed model.

3.1. Reference Prompt

In training stage, the reference prompt comprises speech and its speaking environments, including background sounds and reverberation. The audio encoder of CLAP model processes the reference prompt to obtain the representation of acoustic scenes. Previous studies on AudioLDM and VoiceLDM have demonstrated the capability of generating audio using CLAP embeddings, suggesting their effectiveness in capturing both the acoustic environments and speaking styles of the audio prompts. In the inference stage, we utilize text descriptions of the target acoustic scenes in conjunction with the text encoder of CLAP model, in addition to the reference audio prompt used during training. Since CLAP model is pre-trained using cross-modal self-supervision, we expect that the text description can effectively replace the audio prompt. Utilizing a text prompt offers the potential for greater diversity and enriched representations of acoustic scenes, enabling the model to generate scenarios that have not been encountered before. In our experiments, we simulate reference prompts as noisy, reverberant, or both noisy and reverberant scenes, and also clean speech signals to be transferred into a clean environment.

3.2. Content Prompt

The content prompt consists of the target spoken term that is intended to be retained in outputs of AST-LDM. Since our goal is to transfer acoustic scenes across diverse environments, the content prompt may contain environmental sounds behind speech, and these acoustic elements should not remain in output signals. Consequently, it is essential to precisely preserve the speech signal from the content prompt and condition it onto the diffusion model. To achieve this, we designed a content encoder based on a transformer structure, consisting of two cascaded modules. The content prompt goes into the filter module, which returns a ratio mask using a sigmoid function to be multiplied by the mel-spectrogram of the content prompt. The filter module removes undesired acoustic components, ensuring stable conditioning. Subsequently, the masked mel-spectrogram is passed to the embedding module to encode the content prompts onto the latent space. Like the reference speech, we simulate the content prompt using noisy, reverberant, both noisy and reverberant, and clean speech signals.

3.3. Latent Diffusion

Given the reference and content embeddings as conditions, the latent diffusion model is trained to predict the target audio embedding in the latent domain. As mentioned, the latent embedding of the target audio is extracted using the pre-trained VAE model, which compresses mel-spectrograms into stochastic representations. We utilized the same U-Net structure used in AudioLDM for the diffusion model. The reference embedding is repeated along the time axis and concatenated with the input of the diffusion model, while the content embedding is conditioned using the cross-attention method on the bottleneck feature of the U-Net. The target audio is simulated by utilizing the speech signal of the content prompt and the acoustic scenes of the reference prompt. During the training stage, we used the same speech signal in both reference and content prompts but simulated with different acoustic scenes, and the simulated reference prompt also served as the target audio signal.

4. Experimental Settings

Data Preparation. For training, we utilized 960 hours of speech data from LibriSpeech [30] corpus and curated additional audio samples from AudioCaps [31], a balanced subset of AudioSet [32], and VGGSound [33] to provide background sounds. Room acoustics were simulated using room impulse responses (RIRs) from DNS-Challenge dataset [34], consisting of both simulated and real-recorded samples. We divided the RIR dataset into two subsets: two simulated room impulse responses from each room size and RWCP real impulse responses for evaluation, with the remaining data allocated for training.

For evaluation, we employed both audio samples and text descriptions for the reference prompt. Audio samples were obtained from the *test-clean* subset of the LibriSpeech corpus for speech utterances, the *eval* set of AudioSet, and the RIR evaluation set of DNS-Challenge dataset. We prepared 1,800 speech samples from LibriSpeech, half for clean content and the rest for clean reference prompts. Since we used same speech signal for reference and content prompts during training, we aligned the genders of speakers within reference prompts and content prompts. We overlaid acoustic scenes onto the clean speech samples to create content and reference prompts. In the case of using text descriptions for reference prompts, we generated 1,200 sentences[1] using GPT4 [35], each describing an acoustic scene with the format: *'A [male/female] speaks in [place] with [background sounds] behind.'* For noiseless scenes, we used the text prompt: *'A [male/female] speaks in a quiet room'*.

In addition, to compare with AudioLDM and VoiceLDM, we leveraged audio samples of LibriSpeech and AudioCaps. Since AudioCaps involves text captions of audio samples, we used them as a reference text prompt, as well as a reference audio prompt, and transcriptions of LibriSpeech, were used for the text input of VoiceLDM while we used corresponding speech signals for AST-LDM. We also included real-world recordings from VCTK-DDS dataset, featuring clean-noisy speech pairs captured in real environments with ambient noise and reverberation. For VCTK-DDS dataset [36], we retained audio samples without adding additional background sounds or reverberation, using them as target audio examples. Following the evaluation data ontology of DiffRENT, we selected 2 speakers (p234, p241) and a recording environment, *livingroom1 in Uber*. To ensure consistency across datasets, audio samples were resampled to 16kHz and segmented into 10-second chunks, with shorter samples padded to fit the duration. For the simulation, we mixed speech and environmental sound with a random signal-to-noise ratio (SNR) ranging from 4 to 20 dB.

Model Structure. The architectural configurations of the diffusion model, CLAP model, VAE model, and HiFi-GAN remain consistent with those described in [15, 19]. For the content encoder, we designed it with 2 transformer layers for the filter module and 2 convolution layers followed by 4 transformer layers for the embedding module. Each transformer layer utilizes 256 hidden dimensions with 8 attention heads. Reference prompts, content prompts, and target speech signals are transformed into 64-dimensional mel-spectrograms on a logarithmic scale, extracted every 10ms with a frame length of 64ms.

Training Strategy. We adopt a training scheme similar to that of AudioLDM and VoiceLDM and introduce the dual classifier-free guidance method [29] to train our model conditioned by two different prompts. During training, we randomly dropped reference and content embeddings with probabilities of 0.1 and 0.1, respectively. The model is trained using L2 loss to estimate added noise during the diffusion forward process. While the VAE, CLAP, and HiFi-GAN are pre-trained and left frozen, the diffusion model and content encoder are jointly trained.

Evaluation Protocols. During inference, we used a DDIM sampler with 100 sampling steps, setting 1.0 for both dual classifier-free guidance strengths. We measured performance in terms of scene similarity and content preservation. Scene similarity was assessed using CLAP-based Fréchet Audio Distance (FAD) and CLAP similarity metrics. CLAP similarity was calculated using both the audio encoder ($\text{CLAP}_{\mathcal{A}}$) and, if available, the text encoder ($\text{CLAP}_{\mathcal{T}}$). Content preservation was evaluated using word-error-rate (WER) employing Whisper-small.en [37] model. Also, speaker similarity (SSM) between content prompts and output signals was measured using the embedding distance of RawNet3 [38] model. For the subjective evaluation, we conducted listening tests with 17 listeners to assess the overall perceptual quality of the generated audio samples. The assessment of audio quality was conducted using a 5-point mean-opinion-score (MOS) scale, where scores ranged from 1 (bad) to 5 (excellent). Lower scores for FAD and WER indicate better performance, while higher scores for CLAP, SSM and MOS suggest better performance.

[1] Samples available: `https://ast-ldm.github.io/demo`

Table 1: *Objective evaluation of generated speech with simulated dataset conditioned on audio reference prompts ($\mathcal{R}$: $\mathcal{A}$).*

Scenario ($\mathcal{R}:\mathcal{A}$)	**FAD**	**CLAP**$_{\mathcal{A}}$	**WER** (%)	**SSM**
Clean → Clean	0.345	0.633	4.8	0.814
Clean → Env	0.278	0.627	5.8	0.785
Env → Clean	0.554	0.479	22.6	0.718
Env → Env	0.276	0.609	24.0	0.706

Table 2: *Objective evaluation of generated speech with simulated dataset conditioned on text reference prompts ($\mathcal{R}$: $\mathcal{T}$).*

Scenario ($\mathcal{R}:\mathcal{T}$)	**FAD**	**CLAP**$_{\mathcal{T}}$	**WER** (%)	**SSM**
Clean → Clean	1.096	0.445	6.8	0.779
Clean → Env	0.681	0.463	7.7	0.701
Env → Clean	1.114	0.437	31.3	0.654
Env → Env	0.690	0.446	25.7	0.673

5. Experiment Results

In this section, we present the results of our experiments across different datasets and metrics. Demo audio samples from our experiments are available online[1].

5.1. Performance on Simulated Recordings

In Table 1 and 2, we report the acoustic scene transfer performances of AST-LDM across diverse pairs of content and reference prompts. Table 1 shows AST results using audio signals as reference prompts ($\mathcal{R}:\mathcal{A}$), while Table 2 utilizes text descriptions generated using GPT4 for the prompt ($\mathcal{R}:\mathcal{T}$). Each table showcases 4 scenarios for transferring acoustic scenes: Clean → Clean, Clean → Env, Env → Clean, and Env → Env. Here, 'Env' involves environmental scenes overlaid on speech.

Our findings in Table 1 demonstrate that AST-LDM effectively transfers acoustic scenes while maintaining spoken contents and their speakers. However, AST-LDM appears to be particularly susceptible to the 'Env → Clean' transition. This vulnerability arises from the challenge of describing a 'clean' scene, where CLAP model may not be sufficiently trained on 'silence' or 'quiet' acoustic environments. In Table 2, we found similar results to those obtained using audio reference prompts. Since the generated transcriptions contain complex descriptions and diverse unseen sounding objects, the results show low overall scores compared to Table 1. While it still preserves the spoken term of content prompts well, it encounters difficulties in generating the same environment described in text prompts. We anticipate that these issues in each table can be addressed by fine-tuning CLAP on such clean environmental datasets and utilizing more diverse text descriptions. Additionally, providing suitable instructions or prompts to generate desirable signals through text descriptions is important.

5.2. Performance on Real Recordings

In Table 3, we conducted a comparison between AST-LDM and CLAP-based baseline models, utilizing both audio $\mathcal{A}$ and text captions $\mathcal{T}$ from the AudioCaps dataset as reference prompts $\mathcal{R}$. Based on the FAD and CLAP scores, the proposed model demonstrates competitive generation performance compared to the baseline models. Specifically, the signals generated by AST-LDM exhibit a strong correlation with the textual descriptions (CLAP$_{\mathcal{T}}$) of the desired acoustic scenes. When we measured CLAP scores between audio and text captions of the AudioCaps dataset, they scored 0.521, and the overall CLAP score of our model is the closest to the score. Additionally, there appears to be less performance degradation when using text reference prompts compared to the baseline models. Moreover,

Table 3: *Comparison of baselines and the proposed model. $\mathcal{R}$ denotes the reference prompt.*

Model	$\mathcal{R}$	**FAD**	**CLAP**$_{\mathcal{A}}$	**CLAP**$_{\mathcal{T}}$	**WER** (%)
AudioLDM	$\mathcal{A}$	0.298	0.600	0.349	-
	$\mathcal{T}$	0.697	0.359	0.349	-
	Mean	0.498	0.480	0.349	-
VoiceLDM	$\mathcal{A}$	**0.288**	**0.643**	0.376	14.3
	$\mathcal{T}$	0.753	0.417	0.361	12.2
	Mean	0.521	0.530	0.369	13.3
AST-LDM	$\mathcal{A}$	0.323	0.591	**0.408**	**8.6**
	$\mathcal{T}$	**0.602**	**0.479**	**0.447**	**11.4**
	Mean	**0.463**	**0.535**	**0.428**	**10.0**

Table 4: *Evaluation of generated speech with DDS dataset with environment reference prompts. $\mathcal{C}$ denotes the content prompt. Ground-truth MOS score is 4.206.*

Model	$\mathcal{C}$	**FAD**	**CLAP**$_{\mathcal{A}}$	**SSM**	**MOS**
DiffRENT	Clean	0.332	0.852	0.746	4.000
	Env	0.457	0.792	0.779	4.235
AST-LDM	Clean	0.358	0.842	0.723	3.676
	Env	0.478	0.817	0.772	4.294

our model achieves better WER scores in both types of reference prompts, indicating that AST-LDM effectively preserves the spoken term and accurately conveys it in the output signal. These findings suggest that AST-LDM is capable of generating speech signals accurately within a desired acoustic environment, making it well-suited for the acoustic scene transfer task.

Table 4 presents the evaluation results on DDS datasets compared with DiffRENT model, using the audio samples obtained from its demo page. While both models aim to achieve the same task, DiffRENT is trained exclusively on DDS datasets, which results in a better fit to the evaluation dataset. However, our proposed model still demonstrates competitive performance, and in some metrics, it even outperforms DiffRENT. Besides, AST-LDM has superiority to learn more various audio characteristics using CLAP encoders while DiffRENT is effective in describing simple environments such as ambient noise or reverberation relying on the environmental encoder. Also, with the help of CLAP model, AST-LDM has advantages in describing in ordinary language instruction while DiffRENT always require the target audio sample.

6. Conclusion

In this study, we introduced a novel task, acoustic scene transfer (AST), aimed at transforming the acoustic environments of input speech into target acoustic scenes. We proposed the first AST task-oriented model, AST-LDM, based on the latent diffusion model. By leveraging CLAP model for acoustic scene analysis, AST-LDM successfully altered and represented environments within input speech signals, effectively removing original acoustic scenes. Experimental results have confirmed the potential of AST-LDM for the AST task, demonstrating its ability to transfer acoustic scenes as intended while preserving spoken terms and voice identity in the input speech.

We believe this research would contribute significantly to the field of speech processing, offering enriched and immersive speech generation capabilities. For future work, we recommend focusing on more precise model implementations and refining training strategies. Additionally, we need to explore optimal prompt provisions for generalization and enable wider and more nuanced representations of reference prompts.

7. References

[1] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *INTERSPEECH*, 2018.

[2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*, 2021.

[3] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*, 2020.

[4] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[5] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *ICLR*, 2020.

[6] J. Lee, J. Son Chung, and S.-W. Chung, "Imaginary voice: Face-styled diffusion model for text-to-speech," in *ICASSP*, 2023.

[7] K. Shen, Z. Ju, X. Tan, E. Liu, Y. Leng, L. He, T. Qin, sheng zhao, and J. Bian, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *ICLR*, 2024.

[8] C. Hawthorne, I. Simon, A. Roberts, N. Zeghidour, J. Gardner, E. Manilow, and J. Engel, "Multi-instrument music synthesis with spectrogram diffusion," in *ISMIR*, 2022.

[9] K. Chen, Y. Wu, H. Liu, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies," in *ICASSP*, 2024.

[10] Z. Novack, J. McAuley, T. Berg-Kirkpatrick, and N. J. Bryan, "DITTO: Diffusion Inference-Time T-Optimization for Music Generation," in *ICML*, 2024.

[11] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," in *ICLR*, 2022.

[12] Y. Yuan, H. Liu, X. Liu, X. Kang, M. D. Plumbley, and W. Wang, "Latent diffusion model based foley sound generation system for dcase challenge 2023 task 7," University of Surrey, Guildford, United Kingdom, Tech. Rep., June 2023.

[13] S. Luo, C. Yan, C. Hu, and H. Zhao, "Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models," in *NeurIPS*, 2024.

[14] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," in *ICML*, 2024.

[15] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: text-to-audio generation with latent diffusion models," in *ICML*, 2023.

[16] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[17] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP*, 2023.

[18] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP*, 2023.

[19] Y. Lee, I. Yeon, J. Nam, and J. S. Chung, "Voiceldm: Text-to-speech with environmental context," in *ICASSP*, 2024.

[20] J. Im and J. Nam, "Diffrent: A diffusion model for recording environment transfer of speech," in *ICASSP*, 2024.

[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[22] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *ICLR*, 2020.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[24] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," in *ICLR*, 2020.

[25] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *INTERSPEECH*, 2022.

[26] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP*, 2022.

[27] R. Scheibler, Y. Ji, S.-W. Chung, J. Byun, S. Choe, and M.-S. Choi, "Diffusion-based generative speech source separation," in *ICASSP*, 2023.

[28] S. Lutati, E. Nachmani, and L. Wolf, "Separate and diffuse: Using a pretrained diffusion model for better source separation," in *ICLR*, 2023.

[29] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[31] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *NAACL-HLT*, 2019.

[32] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[33] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020.

[34] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," in *INTERSPEECH*, 2021.

[35] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," Tech. Rep., March 2023.

[36] H. Li and J. Yamagishi, "Dds: A new device-degraded speech dataset for speech enhancement," in *INTERSPEECH*, 2022.

[37] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.

[38] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," in *INTERSPEECH*, 2022.