VOICELDM: TEXT-TO-SPEECH WITH ENVIRONMENTAL CONTEXT

Yeonghyeon Lee, Inmo Yeon, Juhan Nam, Joon Son Chung

Korea Advanced Institute of Science and Technology, South Korea

ABSTRACT

This paper presents VoiceLDM, a model designed to produce audio that accurately follows two distinct natural language text prompts: the description prompt and the content prompt. The former provides information about the overall environmental context of the audio, while the latter conveys the linguistic content. To achieve this, we adopt a text-to-audio (TTA) model based on latent diffusion models and extend its functionality to incorporate an additional content prompt as a conditional input. By utilizing pretrained contrastive language-audio pretraining (CLAP) and Whisper, VoiceLDM is trained on large amounts of real-world audio without manual annotations or transcriptions. Additionally, we employ dual classifier-free guidance to further enhance the controllability of VoiceLDM. Experimental results demonstrate that VoiceLDM is capable of generating plausible audio that aligns well with both input conditions, even surpassing the speech intelligibility of the ground truth audio on the AudioCaps test set. Furthermore, we explore the text-to-speech (TTS) and zero-shot text-to-audio capabilities of VoiceLDM and show that it achieves competitive results. Demos and code are available at https://voiceldm.github.io.

***Index Terms*—** text-to-speech, text-to-audio, latent diffusion model, style control

1. INTRODUCTION

Recent advances in text-to-audio (TTA) generation have shown impressive performance in terms of fidelity and diversity [1, 2, 3, 4, 5, 6, 7]. These models demonstrate the ability to synthesize audio that accurately reflects the semantic context provided by a natural language prompt. Nevertheless, one limitation of these models is that when prompted to produce speech (e.g. *"a man is speaking in a cathedral"*), instead of generating audio with coherent linguistic output, they often generate incoherent babbling voices.

Motivated by this, we introduce VoiceLDM, a text-to-speech (TTS) model inspired by TTA models that also generate linguistically intelligible voices. VoiceLDM can be controlled with two types of natural language prompts, a content prompt specifying the linguistic content of the spoken utterance, and a description prompt that characterizes the environmental context of the audio. Our work can be seen as standing at the intersection of text-to-speech and text-to-audio. To the best of our knowledge, it is the first work that simultaneously achieves the speech intelligibility present in TTS models while also having the diverse audio generation capability found in TTA models. As a result, our model is capable of generating a wide range of sounds, such as speech with sound effects, singing voices, whispering, and more.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00222383).

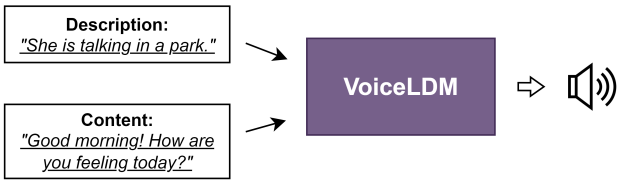

Fig. 1. VoiceLDM produces audio that follows both the description prompt and the content prompt, bridging the gap between the domains of text-to-speech and text-to-audio.

There have been recent or concurrent works in TTS which also possesses the capability to utilize a second text prompt to control the style of the audio being generated [8, 9, 10, 11, 12]. However, the controllable diversity is only limited to speech-related factors such as gender, emotion, and volume.

We build upon the work of AudioLDM [1], a TTA system based on latent diffusion models. We extend the model by integrating an additional content prompt as a conditional input. We train our model using real-world audio data by taking advantage of contrastive language-audio pretraining (CLAP) [13] and Whisper [14]. We are thereby able to use large-scale audio datasets without human annotation, which are then used for model training to achieve better generation results.

Experimental results demonstrate that VoiceLDM generates audio that aligns well with both the content prompt and the description prompt. Furthermore, the audio generated by VoiceLDM often surpasses the linguistic intelligibility of the ground truth audio. We also show that the model is capable of functioning as a regular TTS or TTA model and demonstrate that it achieves competitive results on each task.

2. METHOD

2.1. Model Overview

Figure 2 illustrates the overall framework of VoiceLDM. Given two natural language text prompts $text_{cont}$ and $text_{desc}$, the role of VoiceLDM is to generate audio $\mathbf{X}$ that follows both conditions as input. The description prompt $text_{desc}$ is first converted into a 512-dimensional vector $\mathbf{c}_{desc} \in \mathbb{R}^{512}$ by the pre-trained CLAP [13] model. A reference audio may also be used to attain $\mathbf{c}_{desc}$, since CLAP is designed to project both modalities into the same latent space. The content prompt $text_{cont}$ is encoded into a hidden sequence $\mathbf{H}_{cont} \in \mathbb{R}^{L \times D}$ by the content encoder, where L is the sequence length and D is the dimension size. The differentiable durator then upsamples the hidden sequence into $\mathbf{c}_{cont} \in \mathbb{R}^{N \times D}$, where $L \leq N$. The differentiable durator is identical to the one used in [15]. The U-Net backbone [16] parameterized as θ takes in both

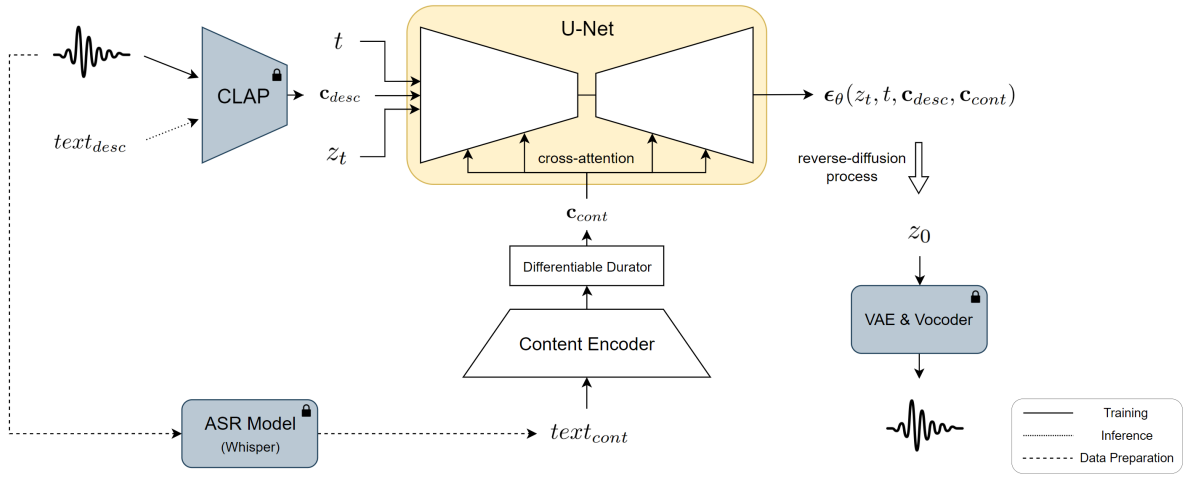

Fig. 2. Overview of VoiceLDM. VoiceLDM is trained with large amounts of real-world audio data. $text_{cont}$ is generated during data preparation by processing the audio with Whisper, an automatic speech recognition (ASR) model. $text_{desc}$ is only used during inference. Modules with a lock icon indicates that it is frozen during training.

conditions $\mathbf{c}_{desc}$ and $\mathbf{c}_{cont}$ and the timestep embedding to predict the diffusion score $\boldsymbol{\epsilon}_\theta$.

Starting from a random noise sampled from an isotropic Gaussian distribution $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the reverse diffusion process iteratively denoises z_t for each time step t with the predicted diffusion score $\boldsymbol{\epsilon}_\theta$ and predicts the initial audio prior z_0. z_0 can then be decoded back to the corresponding mel-spectrogram by the pre-trained variational autoencoder (VAE). Finally, the pre-trained HiFi-GAN vocoder [17] converts the mel-spectrogram into desired audio $\mathbf{X}$.

2.2. Training

The training procedure of VoiceLDM mostly follows the latent diffusion model training procedure as done in [1, 18, 19]. However, the main difference is that the diffusion model utilizes two conditions. Starting from an audio $\mathbf{X}$, the pre-trained VAE compresses the audio into the latent representation z_0. A noisy representation of z_0 at a certain timestep z_t is obtained by applying noise to z_0 through the forward diffusion process, following a predefined noise schedule.

Due to CLAP, manually annotated description prompt $text_{desc}$ is not necessary to obtain $\mathbf{c}_{desc}$ during training. Instead, CLAP is able to take in the original audio $\mathbf{x}$ to obtain the descriptive condition $\mathbf{c}_{desc}$. The content encoder and the differentiable durator encodes the speech transcription $text_{cont}$ into the content condition $\mathbf{c}_{cont}$. Finally, the model is trained to predict the added noise $\boldsymbol{\epsilon}$ with the following re-weighted training objective:

$$\mathcal{L}_\theta = \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(z_t, t, \mathbf{c}_{desc}, \mathbf{c}_{cont})\|_2^2 \tag{1}$$

Parameters of the U-Net backbone, the content encoder and the differentiable durator are all jointly trained. The pre-trained CLAP model, the pre-trained VAE and vocoder are kept frozen during training.

2.3. Dual Classifier-Free Guidance

An interesting property of VoiceLDM is that classifier-free guidance [20] for the reverse diffusion process can be applied independently with respect to each condition $\mathbf{c}_{desc}$ and $\mathbf{c}_{cont}$. This allows one to trade-off mode coverage and sample fidelity for each individual conditions, allowing increased levels of controllability during generation [21]. When two conditions ($\mathbf{c}_{desc}$ and $\mathbf{c}_{cont}$) are viewed as one unified condition, it is possible to apply classifier-free guidance as follows:

$$\begin{aligned}\tilde{\boldsymbol{\epsilon}}_\theta(z_t, \mathbf{c}_{desc}, \mathbf{c}_{cont}) &= \boldsymbol{\epsilon}_\theta(z_t, \mathbf{c}_{desc}, \mathbf{c}_{cont}) \\ &\quad + w\Big(\boldsymbol{\epsilon}_\theta(z_t, \mathbf{c}_{desc}, \mathbf{c}_{cont}) - \boldsymbol{\epsilon}_\theta(z_t, \emptyset)\Big)\end{aligned} \tag{2}$$

where w is the guidance strength and $\emptyset$ indicates the null condition. However, additional control can be achieved by applying dual classifier-free guidance. In this case, the diffusion score $\tilde{\boldsymbol{\epsilon}}$ is formulated as follows:

$$\begin{aligned}\tilde{\boldsymbol{\epsilon}}_\theta(z_t, \mathbf{c}_{desc}, \mathbf{c}_{cont}) &= \boldsymbol{\epsilon}_\theta(z_t, \mathbf{c}_{desc}, \mathbf{c}_{cont}) \\ &\quad + w_{desc}\Big(\boldsymbol{\epsilon}_\theta(z_t, \mathbf{c}_{desc}, \emptyset_{cont}) - \boldsymbol{\epsilon}_\theta(z_t, \emptyset_{desc}, \emptyset_{cont})\Big) \\ &\quad + w_{cont}\Big(\boldsymbol{\epsilon}_\theta(z_t, \emptyset_{desc}, \mathbf{c}_{cont}) - \boldsymbol{\epsilon}_\theta(z_t, \emptyset_{desc}, \emptyset_{cont})\Big)\end{aligned} \tag{3}$$

Derivations for Equation 3 are included in the Appendix[1]. When $w_{desc} = w_{cont}$, its effect is equivalent with that of Equation 2. By appropriately manipulating the values of w_{desc} and w_{cont}, one can effectively regulate the guidance strength for each individual condition. As an example, one may increase the value of w_{cont} but assign a lower value of w_{desc} to obtain audio with increased style diversity while having more linguistic accuracy. An analysis exploring the effect of dual classifier-free guidance is conducted in Section 4.4. To enable the use of dual classifier-free guidance during inference, we randomly drop the conditions $\mathbf{c}_{desc}$, $\mathbf{c}_{cont}$ independently during training.

3. EXPERIMENT SETTINGS

3.1. Data Preparation

We use the following publicly available real-world audio datasets for training: AudioSet [22], the English subset of the CommonVoice 13.0 corpus [23], VoxCeleb1 [24], and DEMAND [25]. To prepare the training dataset, we allocate each audio from these real-world audio datasets into either English speech segments or non-speech

[1] https://voiceldm.github.io

segments. We include all audios from CommonVoice and VoxCeleb as speech segments and include all audios from DEMAND as non-speech segments.

To process AudioSet, we leverage an automatic speech recognition model Whisper [14], where we use two versions of the model: *large-v2* and *medium.en*. *large-v2* is a multilingual model that also has language identification capabilities, whereas *medium.en* is more specialized in English. First we feed all audio into *medium.en* and generate the transcriptions. With the transcriptions from *medium.en*, we classify audio that contains intelligible English speech from those that do not. To further ensure that the audios are correctly classified, we additionally use *large-v2* for audios that have been classified as speech segments. With the language identification functionality of *large-v2*, for each audio we compute the probability of the language being English and generate the transcriptions. We only classify audio as English speech segments if the probability that the language is English is greater than 50%, and the word error rate (WER) between the transcriptions of *large-v2* and *medium.en* is less than 50%.

After the audios are classified into speech segments and non-speech segments, we use the transcriptions generated by *medium.en* to be used as $text_{cont}$ for every audio in the speech segments. We use the transcriptions from *medium.en* instead of *large-v2* since we find that it generates slightly more accurate transcriptions for general audio such as AudioSet. For audios longer than 10 seconds, we take the first 10 seconds of audio before feeding into *medium.en* to generate the transcriptions. For audios that already have pre-existing transcriptions and has a duration of less than 10 seconds, we use the provided transcriptions to achieve better performance.

In total, 2.43M speech segments and 824k non-speech segments are collected. All audio files are resampled into 16kHz sampling rate and mono format. All audios in the speech segments are standardized to have a duration of 10 seconds, either by selecting the initial 10 seconds for longer clips or zero-padding shorter segments.

3.2. Model Configuration

We train two models, VoiceLDM-S and VoiceLDM-M. The difference between these two models is the size of the U-Net backbone. We use the U-Net used in [1], where the channel dimensions of the encoder blocks are $[c_u, 2c_u, 3c_u, 5c_u]$, where c_u is the basic channel number. We use $c_u = 128$ for VoiceLDM-S and $c_u = 192$ for VoiceLDM-M. This results in a total of 280M and 508M number of trainable parameters, including the content encoder and the differentiable durator. To condition the U-Net backbone with two conditions, we replace the self-attention component of the U-Net with cross-attention to additionally condition $\mathbf{c}_{cont}$. $\mathbf{c}_{desc}$ is conditioned in the same way as [1], by concatenating it with the timestep embedding.

We employ the pre-trained VAE and vocoder from [1]. We use the pre-trained CLAP model [13] released by the authors[2]. For the content encoder, it is possible to train a Transformer encoder from scratch. However, we extract a Transformer encoder component from a pre-trained SpeechT5 [26] model trained for TTS[3], in pursuit of improved performance.

3.3. Training Configuration

We use two NVIDIA A5000 GPUs with a batch size of 8 each for VoiceLDM-S and use four NVIDIA A5000 GPUs with a batch size of 4 each to train VoiceLDM-M. Both models are trained for 3M steps. The learning rate is set to $2e-5$ for the AdamW optimizer.

We use audios from the speech segments to train VoiceLDM. However, if the speech segment is from CommonVoice, randomly selected audio from the non-speech segments is mixed on-the-fly with a probability of 0.5. For non-speech segments, the audio is randomly cut or padded to have a duration of 10 seconds and is mixed with a signal-to-noise ratio (SNR) value randomly selected from a uniform distribution within the range of $[4,\ 20]$. Otherwise if the speech segment is from AudioSet or VoxCeleb, we do not mix non-speech audio since the audio is already sufficiently noisy.

During training, $\mathbf{c}_{desc}$ and $\mathbf{c}_{cont}$ are randomly dropped with a probability of 0.1 respectively. During inference, we use a DDIM sampler [27] with 100 as the number of inference steps.

Table 1. Performance comparison with quantitative metrics on the AudioCaps test set. ↑: higher is better; ↓: lower is better.

Model	FAD↓	KL↓	CLAP↑	WER(%)↓	ΔWER(%)↓
Ground Truth	-	-	0.251	21.21	21.21
VoiceLDM-S	4.781	1.454	0.210	56.03	47.56
VoiceLDM-M	5.62	1.48	0.197	13.05	13.22
VoiceLDM-M$_{audio}$	2.499	0.883	0.209	13.41	11.82
AudioLDM 2	20.720	3.005	0.060	32.84	27.39

3.4. Evaluation Metrics

We use quantitative and qualitative metrics to assess the audio quality and the input prompt adherence of VoiceLDM.

Quantitative Metrics. We report Frechet Audio Distance (FAD), Kullback-Leiber (KL) divergence, and CLAP score. Additionally, to evaluate speech intelligibility, we measure the word error rate (WER) with Whisper *large-v2*. We also report the word error rate (ΔWER) between the transcriptions of two Whisper models, *large-v2* and *medium.en*. Having a lower value of ΔWER suggests that the generated audio has high speech intelligibility.

Qualitative Metrics. We report overall impression (OVL), relevance between audio and condition (REL), and mean opinion score (MOS) of the generated audio. For qualitative evaluation, we use crowd-sourcing and ask participants to rate the audio on a scale between 1 to 5. We make sure each audio is evaluated by at least 10 different raters.

4. RESULTS

4.1. Main Result

We evaluate the performance of VoiceLDM on the AudioCaps [28] test set. Segments containing English speech are collected and the corresponding transcriptions are generated as described in Section 3.1. We denote the original test set as *ac-full* and the processed test set as *ac-filtered*. We use the captions from AudioSet as $text_{desc}$ and the generated transcriptions as $text_{cont}$. We use $w_{desc} = 7, w_{cont} = 7$ for dual classifier-free guidance. We also substitute $text_{desc}$ with the ground truth audio for the descriptive condition, and denote the experiment setting as VoiceLDM-M$_{audio}$. For objective evaluation, we additionally compare VoiceLDM with an AudioLDM 2 [8] checkpoint trained for TTS[4], a model also capable of accepting a description and content prompt to generate audio.

Quantitative and qualitative evaluation results are shown in Table 1 and Table 2. VoiceLDM is capable of generating audio that

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

[3]https://huggingface.co/microsoft/speecht5_tts

[4]We use the *audioldm2-speech-gigaspeech* checkpoint.

Table 2. Performance comparison with qualitative metrics on the AudioCaps test set. We report overall quality (OVL), relevance between the audio and descriptive prompt (REL$_{desc}$), and the relevance between the audio and content prompt (REL$_{cont}$).

Model	OVL	REL$_{desc}$	REL$_{cont}$
Ground Truth	4.27	4.30	4.45
VoiceLDM-S	3.61	3.66	4.14
VoiceLDM-M	3.88	3.89	4.52
VoiceLDM-M$_{audio}$	4.08	4.03	4.61

Table 3. Performance comparison on TTS capabilities on the CommonVoice test set.

Model	WER(%)↓	ΔWER(%)↓	MOS
Ground Truth	11.818	11.818	4.09
VoiceLDM-S	10.693	8.378	3.74
VoiceLDM-M	3.909	2.390	3.96
VoiceLDM-M$_{audio}$	10.459	7.911	3.89
FastSpeech 2 [29]	9.751	9.327	3.32
SpeechT5 [26]	6.384	3.054	3.74

adheres to both input conditions simultaneously. The largest model VoiceLDM-M, even surpasses the speech intelligibility of the ground truth audio, while maintaining competitive audio quality and description prompt adherence. Substituting $text_{desc}$ with the audio yields improved outcomes in following the environmental context, while also achieving high speech intelligibility. AudioLDM 2, a TTS-focused model, fails to adhere to the given description prompt if the prompt encompasses more than just speech-related elements.

4.2. Text-to-Speech Capabilities

VoiceLDM has the ability to act as a regular TTS model with the prompt *"clean speech"* as input for $text_{desc}$. We evaluate the TTS capabilities of VoiceLDM on the CommonVoice test set. The transcriptions from the CommonVoice test set are given as $text_{cont}$. We use $w_{desc} = 1, w_{cont} = 9$ for dual classifier guidance. We compare the performance with SpeechT5 trained for TTS and FastSpeech 2 trained on CommonVoice[5].

Table 3 shows the results of TTS evaluation. Evaluation on the CommonVoice test set reveals that all VoiceLDM models are able to surpass the ground truth audio in terms of linguistic intelligibility, as measured by WER and ΔWER. The largest model, VoiceLDM-M achieves the lowest WER and ΔWER and even achieves naturalness comparable to ground truth audio. VoiceLDM-M also outperforms FastSpeech 2 and SpeechT5 across all metrics by a significant margin.

4.3. Text-to-Audio Capabilities

Although VoiceLDM is trained solely on audio samples with human voices, it exhibits the ability to perform regular zero-shot TTA. We evaluate the zero-shot TTA capabilities of VoiceLDM on the AudioCap test set. The captions provided from the AudioCap test set are given as $text_{desc}$, and an empty string is given as $text_{cont}$. $w_{desc} = 9, w_{cont} = 1$ is used for dual classifier guidance.

The results in Table 4 show that despite not being specifically trained for TTA, VoiceLDM is capable of generating plausible audio as seen in TTA models. VoiceLDM-M achieves comparable results

[5]https://huggingface.co/facebook/fastspeech2-en-200_speaker-cv4

Table 4. Performance comparison on TTA capabilities on the AudioCaps test set.

Model	FAD↓	KL↓	CLAP↑
ac-full			
Ground Truth	-	-	0.259
VoiceLDM-S	15.073	3.309	0.089
VoiceLDM-M	10.119	2.458	0.172
AudioLDM-S	5.131	1.823	0.178
AudioLDM-M	4.689	1.986	0.224
ac-filtered			
Ground Truth	-	-	0.247
VoiceLDM-S	9.852	2.504	0.099
VoiceLDM-M	6.819	1.713	0.185
AudioLDM-S	3.211	1.485	0.175
AudioLDM-M	2.974	1.621	0.215

Table 5. Effect of dual classifier-free guidance.

w_{desc}	w_{cont}	FAD↓	KL↓	CLAP↑	WER(%)
5	5	5.08	1.45	0.193	15.95
5	7	5.60	1.53	0.172	13.95
5	9	6.15	1.69	0.170	12.41
7	5	5.25	1.42	0.198	17.67
7	7	5.62	1.48	0.197	13.05
7	9	6.16	1.54	0.175	12.65
9	5	5.50	1.44	0.196	18.46
9	7	5.81	1.47	0.190	13.63
9	9	6.14	1.54	0.177	13.97

in terms of KL and CLAP scores when compared with AudioLDM-S, a model specifically trained for TTA. The gap in performance becomes smaller when evaluated on the *ac-filtered* test set, even outperforming AudioLDM-S in terms of CLAP score.

4.4. Effect of Dual Classifier-Free Guidance

We conduct a series of experiments on the *ac-filtered* test set to explore the effect of dual classifier-free guidance. We compare the performance of VoiceLDM-M by only adjusting the values of w_{desc} and w_{cont}.

As shown in Table 5, while using a high value of w_{cont} yields high speech intelligibility, it leads to a trade-off in reduced adherence to the description prompt. Conversely, increasing w_{desc} enhances adherence but compromises speech intelligibility. Adjusting the value of w_{desc} and w_{cont} allows one to balance this trade-off, thereby facilitating the generation of more desirable outcomes.

5. CONCLUSION

This paper introduces VoiceLDM, a model that introduces unique functionality to control TTS generation with environmental context. VoiceLDM is trained with vast quantities of real-audio data through the utilization of CLAP and Whisper. We improve model controllability by employing dual classifier-free guidance, which enables one to control the trade-off of the guidance strength for each condition. Quantitative and qualitative evaluation results show that VoiceLDM is simultaneously capable of achieving the speech synthesis and general audio synthesis functionalities found in TTS and TTA models. Furthermore, we show that VoiceLDM can function as a conventional TTS or TTA model, positioning itself as a generalized extension of the two domains.

6. REFERENCES

[1] 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 *Proc. ICML*, 2023.

[2] 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 *Proc. ICLR*, 2023.

[3] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *Proc. ICML*, 2023.

[4] J. Huang, Y. Ren, R. Huang, D. Yang, Z. Ye, C. Zhang, J. Liu, X. Yin, Z. Ma, and Z. Zhao, "Make-an-audio 2: Temporal-enhanced text-to-audio generation," *arXiv preprint arXiv:2305.18474*, 2023.

[5] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[6] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: a language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[7] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

[8] H. Liu, Q. Tian, Y. Yuan, X. Liu, X. Mei, Q. Kong, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *arXiv preprint arXiv:2308.05734*, 2023.

[9] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "Prompttts: Controllable text-to-speech with text descriptions," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[10] D. Yang, S. Liu, R. Huang, G. Lei, C. Weng, H. Meng, and D. Yu, "Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt," *arXiv preprint arXiv:2301.13662*, 2023.

[11] G. Liu, Y. Zhang, Y. Lei, Y. Chen, R. Wang, Z. Li, and L. Xie, "Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions," *arXiv preprint arXiv:2305.19522*, 2023.

[12] M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim, and N. S. Kim, "Expressive Text-to-Speech Using Style Tag," in *Proc. Interspeech*, 2021, pp. 4663–4667.

[13] 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 *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*. PMLR, 2023, pp. 28 492–28 518.

[15] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "Naturalspeech: End-to-end text to speech synthesis with human-level quality," *arXiv preprint arXiv:2205.04421*, 2022.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Springer, 2015, pp. 234–241.

[17] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR*, 2022, pp. 10 684–10 695.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, vol. 33, 2020, pp. 6840–6851.

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

[21] W. Cho, H. Ravi, M. Harikumar, V. Khuc, K. K. Singh, J. Lu, D. I. Inouye, and A. Kale, "Towards enhanced controllability of diffusion models," *arXiv preprint arXiv:2302.14368*, 2023.

[22] 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 *Proc. ICASSP*. IEEE, 2017, pp. 776–780.

[23] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[24] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *Telephony*, vol. 3, pp. 33–039, 2017.

[25] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.

[26] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang *et al.*, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proc. ACL*, 2022, pp. 5723–5738.

[27] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. ICLR*, 2020.

[28] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proc. NAACL*, 2019, pp. 119–132.

[29] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.

A. DERIVATION OF DUAL CLASSIFIER-FREE GUIDANCE

Given two conditions c_1, c_2, let $p_\theta(z_t|c_1, c_2)$ be the density of the conditional distribution of z_t, which is estimated by a score prediction network θ. When applying classifier-free guidance [20] for two conditions, the conditional distribution of p_θ is modified with additional guidance with strength w as follows:

$$\tilde{p}_\theta(z_t|c_1, c_2) \propto p_\theta(z_t|c_1, c_2)p_\theta(c_1, c_2|z_t)^w \tag{4}$$

For the case of VoiceLDM, it is reasonable to assume that the two conditions c_1 and c_2 are independent. In this case, the conditional distribution is modified as follows:

$$\tilde{p}_\theta(z_t|c_1, c_2) \propto p_\theta(z_t|c_1, c_2)p_\theta(c_1|z_t)^w p_\theta(c_2|z_t)^w \tag{5}$$

One may also consider the possibility of using different guidance strengths for each condition, where we denote the individual guidance strengths as w_1 and w_2:

$$\tilde{p}_\theta(z_t|c_1, c_2) \propto p_\theta(z_t|c_1, c_2)p_\theta(c_1|z_t)^{w_1} p_\theta(c_2|z_t)^{w_2} \tag{6}$$

From this we get the gradient of the log-density of the modified conditional distribution as

$$\begin{aligned}
&\nabla_{z_t} \log \tilde{p}_\theta(z_t|c_1, c_2) \\
&= \nabla_{z_t} \log p_\theta(z_t|c_1, c_2)p_\theta(c_1|z_t)^{w_1} p_\theta(c_2|z_t)^{w_2} \\
&= \nabla_{z_t} \log p_\theta(z_t|c_1, c_2)\left(\frac{p_\theta(z_t|c_1)}{p_\theta(z_t)}\right)^{w_1}\left(\frac{p_\theta(z_t|c_2)}{p_\theta(z_t)}\right)^{w_2} \\
&= \nabla_{z_t} \log p_\theta(z_t|c_1, c_2) \\
&\quad + w_1\Big(\nabla_{z_t} \log p_\theta(z_t|c_1) - \nabla_{z_t} \log p_\theta(z_t)\Big) \\
&\quad + w_2\Big(\nabla_{z_t} \log p_\theta(z_t|c_2) - \nabla_{z_t} \log p_\theta(z_t)\Big)
\end{aligned} \tag{7}$$

Finally, this can be rewritten in terms of diffusion scores:

$$\begin{aligned}
\tilde{\boldsymbol{\epsilon}}_\theta(z_t, c_1, c_2) = \boldsymbol{\epsilon}_\theta(z_t, c_1, c_2) \\
+ w_1\Big(\boldsymbol{\epsilon}_\theta(z_t, c_1) - \boldsymbol{\epsilon}_\theta(z_t)\Big) \\
+ w_2\Big(\boldsymbol{\epsilon}_\theta(z_t, c_2) - \boldsymbol{\epsilon}_\theta(z_t)\Big)
\end{aligned} \tag{8}$$