Text-to-Audio Generation via Bridging Audio Language Model and Latent Diffusion

Zhenyu Wang[1], **Yong Xu**[2], **Chenxing Li**[2], **ChunLei Zhang**[2], **John H.L. Hansen**[1], **Dong Yu**[2]
[1] The University of Texas at Dallas, USA
[2] Tencent AI Lab, Bellevue, USA
zhenyu.wang@utdallas.edu,lucayongxu@global.tencent.com
lichenxing007@gmail.com,zhangclei89@gmail.com
john.hansen@utdallas.edu,dyu@global.tencent.com

Abstract

Diffusion models have become the foundation for most text-to-audio generation methods. These approaches rely on a large text encoder to process the textual description, serving as a semantic condition to guide the audio generation process. Meanwhile, autoregressive language model-based methods for audio generation have also emerged. These autoregressive models offer flexibility by predicting discrete audio tokens, but they often fail to achieve high fidelity. In this work, we propose an advanced system that integrates the autoregressive language model with the diffusion model, achieving flexible and refined audio generation. The autoregressive language model is used to predict the discrete audio tokens conditioned on text prompts. Then, audio tokens are fed into the diffusion model to further purify the details of the generated audio. Consequently, compared to baseline systems, our proposed approach can deliver better results on most objective and subjective metrics on the AudioCaps test set. Audio demos generated by our proposed best system are available at `https://dcldmdemo.github.io`.

1 Introduction

Following the revolution of the text-to-image (TTI) generation [1, 2, 3], text-to-audio (TTA) models have made significant strides [4, 5, 6]. TTA tasks aim to produce audio content from a text description. Such models hold promising potential for applications such as media production and audio novels. In previous research, the methods of TTA can be broadly categorized into two separate categories: (i) auto-regressive (AR) or non-auto-regressive (NAR) transformer-based models, typically manifested as language models, often working with discrete audio representations. With a multi-scale transformer model, UniAudio [7] utilizes large language model techniques to generate various audio types, including speech, sounds, music, and singing. MAGNet [6] fuses AR and NAR models designed for efficient operation. The representative AR generative model (AUDIOGEN [8]) utilizes learned discrete audio representations with a transformer decoder to generate audio conditioned on textual embeddings. (ii) diffusion-based models, typically functioning with continuous latent representations of the audio signal. Models such as AudioLDM [5], AudioLDM2 [9], Tango [10], Make-an-audio [11], and Make-an-audio2 [12], leverage latent variable generation coupled with pre-trained Variational Auto-encoder (VAE) [13] and HiFi-GAN [14] for audio reconstruction, achieving notable successes.

We integrate the diffusion model with an AR model by leveraging the strengths of both models: the fidelity and noise resilience of the diffusion model and the flexibility of the AR model. Recently, the combination of an AR transformer and diffusion model has achieved better performance on the text-to-speech task [15, 16]. In TTA, the AR model can flexibly predict the discrete audio tokens

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Workshop on AI-Driven Speech, Music, and Sound Generation.

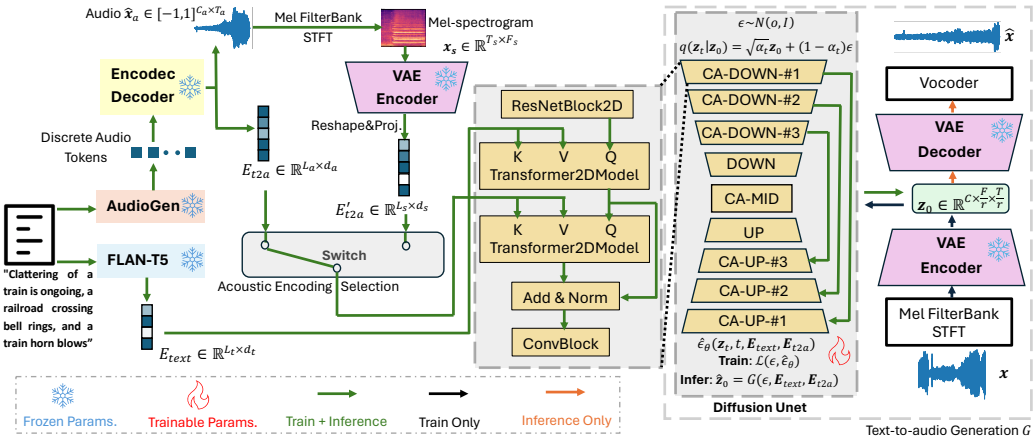


Figure 1: Dual-conditioned latent diffusion model architecture. CA-DOWN denotes cross-attention downsampling blocks, CA-UP denotes cross-attention upsampling blocks, and CA-MID denotes middle cross-attention blocks (the intermediate feature shape is maintained). Parameters in U-Net are updated during training, while other components' parameters are fixed.

conditioned on text prompts. The diffusion model is followed to generate high-fidelity audio with the input of the predicted discrete tokens from the AR model. Different from AudioLDM2 [9], where the AR model is used to obtain semantic tokens, our system gets acoustic tokens from the AR model.

Specifically, we propose a dual-conditioned latent diffusion model (DC-LDM) conditioned both on the text embedding from a large language model (FLAN-T5) and the acoustic representations from an autoregressive audio language model [8]. Our main contributions are listed as follows: **(1)** We proposed an advanced method to generate high-fidelity audio by bridging the AR audio language model and the latent diffusion model. **(2)** We further explored another conditional embedding option of the acoustic latent features, which is compressed via a pre-trained VAE. **(3)** We employed an additional cross-attention block between intermediate features with acoustic encoding to formulate a dual-conditioned diffusion model. **(4)** We evaluated our text-to-audio generation approach on the public AudioCaps test set, allowing for parallel comparison with existing systems. We present objective and subjective metrics demonstrating that our method gains improvement over the evaluated baselines. Additionally, we include an ablation study to elucidate the contributions.

2 Methodology

The proposed DC-LDM is depicted in Fig. 1. Initially, the textual-prompt encoder processes input descriptions to generate textual embedding. Additionally, we use a text-to-audio pre-trained model to generate acoustic latent features as an auxiliary representation. These two representations are added with a skip-connection and then utilized to create a latent audio representation or audio prior from standard Gaussian noise through reverse diffusion. The VAE decoder produces a mel-spectrogram from the latent representation. This mel-spectrogram is fed into a vocoder to synthesize audio outputs.

2.1 Textual-prompt Encoder

We leverage the pre-trained large language model (LLM) FLAN-T5-Large (780M) [17] as the text encoder to facilitate TTA generation. The resulting text embedding is termed as $\mathbf{E}_{text} \in \mathbb{R}^{L_t \times d_t}$, where L_t and d_t are the number of tokens and corresponding embedding size, respectively.

2.2 Acoustic Latent Feature Encoder

AUDIOGEN [8] uses a two-stage process involving a neural audio compression model [18] to encode raw audio into discrete tokens, and an AR transformer-decoder to generate target discrete tokens, conditioned on text inputs. Target discrete tokens can be decoded into continuous latent encoding $\mathbf{E}_{t2a} \in \mathbb{R}^{L_a \times d_a}$ (L_a and d_a are the sequence length and the corresponding embedding size, respectively). Alternatively, tokens can be decoded into a reconstructed representation/audio

signal $\hat{\mathbf{x}}_a \in [-1, 1]^{C_a \times T_a}$ (C_a denotes the number of audio channels, $T_a = d \times f_{sr}$ is the number of audio samples at a given sample rate f_{sr}). This audio can be further processed into the mel-spectrogram $\mathbf{X}_s \in \mathbb{R}^{T_s \times F_s}$. Subsequently, the mel-spectrogram can be compressed by the audio VAE [13] into audio prior $\mathbf{z}_s \in \mathbb{R}^{C \times \frac{F_s}{r} \times \frac{T_s}{r}}$ (see Sec. 2.4), then reshaped as $\mathbf{E}'_{t2a} \in \mathbb{R}^{L_s \times d_s}$, where $L_s = \frac{T_s}{r}, d_s = C \times \frac{F_s}{r}$. Either $\mathbf{E}_{t2a}$ or $\mathbf{E}'_{t2a}$ can be considered as an acoustic encoding to guide audio generation (see Sec. 2.3).

2.3 Dual-conditioned Latent Diffusion Model

The latent diffusion model (LDM) [2] is capable of generating an audio sample $\hat{\mathbf{x}}$ from given conditions (i.e., text description $\mathbf{E}_{text}$, acoustic information $\mathbf{E}_{t2a}$). Using probabilistic generative models like LDMs, we approximate true conditional data distributions $q(\mathbf{z}_0|\mathbf{E}_{text}, \mathbf{E}_{t2a})$ with parameterized $p_\theta(\mathbf{z}_0|\mathbf{E}_{text}, \mathbf{E}_{t2a})$, where $\mathbf{z}_0$ represents the prior of an audio sample within a compressed space of the mel-spectrogram $\mathbf{X}$ (see Sec. 2.4). $\mathbf{E}_{text}$ is text embedding produced by the pre-trained text encoder (see Sec. 2.1), $\mathbf{E}_{t2a}$ is the acoustic encoding generated by the acoustic latent feature encoder (see Sec. 2.2). The LDM in this work is adapted from [19]. The forward transition is a T-step Markov chain process without any trainable parameters. Given the prior $\mathbf{z}_{t-1}$ at diffusion step $t-1$, the data distribution of $\mathbf{z}_t$ at step t can be written as,

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \sqrt{1-\beta_t}\mathbf{z}_{t-1} + \sqrt{\beta_t}\epsilon, \tag{1}$$

where the noise schedule hyper-parameter $0 < \beta_1 < \beta_2 < \cdots < \beta_T < 1$ determines noisier versions of $\mathbf{z}_0$ at each step t. By recursive substitution of $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ in Eq. 1, which allows direct sampling of z_t from z_0 via a non-Markovian process,

$$q(\mathbf{z}_t|\mathbf{z}_0) = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\epsilon, \tag{2}$$

where $\alpha_t = \Pi_{t-1}^t 1 - \beta_t$, $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$. The distribution of z_t will be close to a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, I)$ at the forward process final step $t = T$. The backward transition reconstructs z_0 starting from the Gaussian noise distribution $p(\mathbf{z}_t) \sim \mathcal{N}(\mathbf{0}, I)$ via a dual-conditioned noise estimation ($\hat{\epsilon}_\theta$), the loss function for parameter optimization is formulated as,

$$\mathcal{L} = \sum_{t=1}^{T} \gamma_t \mathbb{E}_{\epsilon_t \sim \mathcal{N}(\mathbf{0}, I), \mathbf{z}_0} ||\epsilon_t - \hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{E}_{text}, \mathbf{E}_{t2a})||_2^2, \tag{3}$$

where γ_t is the weight of reverse step t [20]. The audio prior $\mathbf{z}_0$ is iteratively generated from $\mathbf{z}_t$ conditioned on text embedding $\mathbf{E}_{text}$ and acoustic encoding $\mathbf{E}_{t2a}$:

$$p_\theta(\mathbf{z}_{0:T}|\mathbf{E}_{text}, \mathbf{E}_{t2a}) = p(\mathbf{z}_T)\Pi_{t-1}^T p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{E}_{text}, \mathbf{E}_{t2a}), \tag{4}$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{E}_{text}, \mathbf{E}_{t2a}) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t, \mathbf{E}_{text}, \mathbf{E}_{t2a}), \sigma_t^2 I). \tag{5}$$

The mean and variance are calculated as [19],

$$\mu_\theta(\mathbf{z}_t, t, \mathbf{E}_{text}, \mathbf{E}_{t2a}) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{E}_{text}, \mathbf{E}_{t2a})), \tag{6}$$

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \tag{7}$$

where $\sigma_1^2 = \beta_1$, and the noise estimation $\hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{E}_{text}, \mathbf{E}_{t2a})$ is predicted with U-Net [21], which leverages textual guidance $\mathbf{E}_{text}$ and acoustic guidance $\mathbf{E}_{t2a}$ via cross-attention.

2.4 Audio VAE and Vocoder

The audio VAE [13] compresses the mel-spectrogram of an audio sample $\mathbf{X} \in \mathbb{R}^{T \times F}$ into an audio prior $\mathbf{z}_0 \in \mathbb{R}^{C \times \frac{F}{r} \times \frac{T}{r}}$. Here, r indicates the compression level, C denotes the channel of compressed representations, F and T are frequency and time dimensions in the mel-spectrogram $\mathbf{X}$.

2.5 Classifier Free Guidance

Classifier-free guidance [22] is used during inference, a guidance scale w determines the contribution of text $\mathbf{E}_{text}$ and acoustic $\mathbf{E}_{t2a}$ guidance to noise estimation $\hat{\epsilon}_\theta$, while the empty text is passed to obtain both encodings during unguided estimation:

$$\hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{E}_{text}, \mathbf{E}_{t2a}) = w\epsilon_\theta(\mathbf{z}_t, t, \mathbf{E}_{text}, \mathbf{E}_{t2a}) + (1-w)\epsilon_\theta(\mathbf{z}_t, t). \tag{8}$$

Table 1: Experimental results on the AudioCaps evaluation set. The top section shows results for LDM conditioned only on FLAN-T5-Large encoder output $\mathbf{E}_{text}$ (T5) (see Sec. 2.1)/ acoustic encoding $\mathbf{E}_{t2a}$ converted from AUDIOGEN (AG) generated audio tokens (see Sec. 2.2)/ acoustic encoding $\mathbf{E}'_{t2a}$ converted by VAE (see Sec. 2.4) from AUDIOGEN generated audio (see Sec. 2.2) (AG_VAE). Given T5 embedding $\mathbf{E}_{text}$ as the primary condition, DC-LDM generation performance with either $\mathbf{E}_{t2a}$ or $\mathbf{E}'_{t2a}$ as the auxiliary condition is present in the middle and bottom sections, respectively. CA-DOWN-#1 represents the first cross-attention down-sampling block, CA-UP-#1 denotes the last cross-attention up-sampling block, and CA-MID is the middle cross-attention block in the U-Net (see Fig. 1). #Params. denotes trainable parameters.

Model	#Params.	KL_σ	KL	FAD	FD
LDM w/ T5 (reproduced from TANGO [10])	866M	4.01	1.43	**2.22**	26.86
LDM w/ AG	866M	5.33	1.95	4.80	33.90
LDM w/ AG+VAE	866M	5.85	2.17	4.91	36.21
DC-LDM w/ T5 & AG (CA-DOWN-#1)	873M	3.37	1.28	2.60	24.64
DC-LDM w/ T5 & AG (CA-MID)	899M	3.74	1.38	2.63	22.82
DC-LDM w/ T5 & AG (CA-UP-#1)	875M	3.64	1.34	2.37	22.08
DC-LDM w/ T5 & AG (CA-DOWN#1 & CA-UP#1)	877M	3.30	**1.22**	2.24	**21.86**
DC-LDM w/ T5 & AG_VAE (CA-DOWN#1)	873M	3.89	1.38	2.55	23.60
DC-LDM w/ T5 & AG_VAE (CA-MID)	899M	3.79	1.42	2.74	24.05
DC-LDM w/ T5 & AG_VAE (CA-UP#1)	875M	3.75	1.39	2.55	22.71
DC-LDM w/ T5 & AG_VAE (CA-DOWN#1& CA-UP-#1)	877M	**3.24**	1.23	2.61	22.98

3 Experiment

We employ the AudioCaps dataset [23], which comprises 45222 audio clips in train set. The dataset also includes a validation set and evaluation set with 2,240 instances and 957 instances, respectively.

We adopt both objective and subjective evaluation: **1)** we calculate the Fréchet Audio Distance (FAD) [24]. Similar to FAD, Frechet Distance (FD) [5] utilizes a different classifier, replacing VGGish [25] with PANNs [26]. Additionally, we measure the KL-Divergence (two metrics involved, KL: softmax over logits, KL_σ: sigmoid over logits); **2)** the generated samples are also rated by human based on overall generation quality (OGL), relevance to the input text (REL), and audio quality (AQ) on a scale from 1 to 100. OGL analyzes the semantic consistency between audio and text, the quality of audio, etc. REL measures completeness and sequential consistency. AQ is used to evaluate the quality of generated audio, such as audio clarity and intelligibility. Eight professional annotators are employed, and 100 test audio samples are randomly selected from the AudioCaps test set [23].

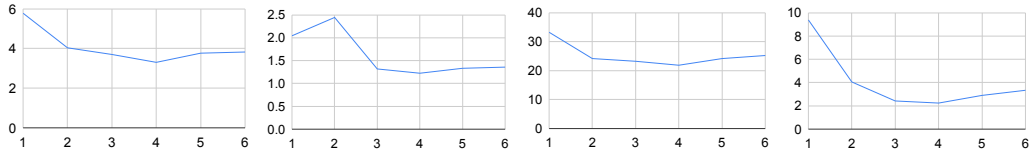
This DC-LDM is based on the Stable Diffusion U-Net architecture [21], applying 8 channels and cross-attention dimensions of 1024 and 128 for text embedding and acoustic encoding, respectively. The acoustic encoding is generated by the AUDIOGEN model [8] with a duration of 10s per audio clip. We employ the AdamW optimizer with a learning rate of 3e-5 for optimization, using a linear learning rate scheduler throughout the training process.

4 Result analysis and discussion

Following the setup in [10, 19], we train the latent diffusion with FLAN-T5 embedding as the condition. As described in Sec. 2.2, we replace the FLAN-T5 embedding $\mathbf{E}_{text}$ with the AUDIOGEN embedding $\mathbf{E}_{t2a}$. Besides, audio tokens can be converted into a raw waveform, which can be further processed by the VAE model, then reshaped and projected into a latent feature $\mathbf{E}'_{t2a}$. We assume this audio prior derived from AUDIOGEN and VAE can be another option as the diffusion auxiliary condition. As shown in Table 1, experimental results do not meet our expectations. As a result of the bias introduced by AUDIOGEN, generated audios might be overfitting to the data distribution derived from AUDIOGEN predictions, which could not be well aligned with the diffusion training objective. We further keep the FLAN-T5 embedding as the primary condition and employ acoustic encoding as an auxiliary condition. On top of the original model setup, we insert a cross-attention layer right after the designated cross-attention layers, refer to Fig. 1) to involve the acoustic encoding in the model training, and we apply skip-connection for previous cross-attention output to maintain the model performance. As depicted in Table 1, two options (i.e. $\mathbf{E}_{t2a}$/$\mathbf{E}'_{t2a}$) for extracted acoustic encoding

Table 2: Performance comparison on the AudioCaps evaluation set with existing systems. AS, AC, SS, P5, and ppd denote AudioSet, AudioCaps, ShutterStock, Pond5, and proprietary data, respectively. #Params. denotes trainable parameters.

Model	Datasets	#Params.	Objective metrics			Human subjective scores		
			KL	FAD	FD	OGL	REL	AQ
Ground-truth	-	-	-	-	-	93.72	93.64	94.45
DiffSound [4]	AS+AC (5,565 hrs)	400M	2.52	7.75	47.68	-	-	-
AudioLDM-L [5]	AC (145 hrs)	739M	1.86	2.08	27.12	90.27	89.04	91.17
AudioLDM2-full [9]	AS+AC+6 others (29,510 hrs)	346M	1.58	3.39	25.75	92.1	89.84	92.27
MAGNET[6]	SS+P5+ppd (20,000 hrs)	1.5B	1.64	2.36	-	-	-	-
AUDIOGEN-Large [8]	AS+AC+8 others (6,824 hrs)	1B	1.69	1.82	-	91.04	90.92	92.61
TANGO [10]	AC (145 hrs)	866M	1.37	**1.59**	24.52	91.89	92.08	**93.94**
Proposed DC-LDM	AC (145 hrs)	877M	**1.22**	2.24	**21.86**	**92.74**	**92.16**	93.63



(a) KL_σ (b) KL (c) FD (d) FAD

Figure 2: The effect of classifier-free guidance scale. Numbers on the x-axis denote different guidance scales. The y-axis represents the numerical scales of each metric.

can help improve the generation performances. Table 1 shows the effect of inserting additional cross-attention modules into different blocks, which is marginally beneficial to system performance improvement while incorporating the additional condition into three different blocks of the U-Net. The best-performing system integrates the additional cross-attention module for intermediate latent embedding $\mathbf{E}_{text}$ and acoustic encoding $\mathbf{E}_{t2a}$ at both CA-DOWN-#1 and CA-UP-#1. The Classifier-free guidance scale represents a trade-off between sample diversity and conditional generation quality. We show the effect of varied guidance scale w on TTA generation performance in Fig. 2. When $w = 4$, we achieve the best results across all evaluation metrics. These numbers under difference guidance scales are obtained from the best-performing system in Table. 1.

In Table 2, we compare our proposed approach with existing systems, including DiffSound [4], AUDIOGEN [8], AudioLDM [5], AudioLDM2 [9], MAGNET[6], and TANGO [10]. Regarding the automatic objective evaluation, DC-LDM achieves superior results on two metrics (i.e., KL and FD), and the proposed system stays comparable on the FAD evaluation. Overall, the proposed DC-LDM model is only trained on the AudioCaps dataset, and it delivers promising scores with 1.22 KL, 2.24 FAD, and 21.86 FD. We suppose the performance gap might be resulting from the less robust audio classifier in FAD compared to FD, as discussed in Sec. 3. Consequently, enhancements in adherence to detailed language descriptions could potentially provide misleading information to the classifier in FAD. Although FAD is conceptually similar to FD, it employs VGGish [24] as its classifier, which might underperform the classifier PANNs used in FD [5]. From Table 2. subjective evaluation results show significant gains of DC-LDM with OGL of 91.74 and REL of 92.16, outperforming existing systems in terms of OGL and REL and maintaining comparable results in AQ evaluation. The subjective results indicate that evaluators prefer our model-generated audio against existing systems in terms of audio naturalness and faithfulness.

5 Conclusion

We introduce a novel DC-LDM, which bridges an AR audio language model and a LDM. Given FLAN-T5 embedding as a primary condition in diffusion models, we additionally use an acoustic latent encoding from the AR model to exploit complementary information in the audio generation process. We use a skip connection to combine the outputs of two cross-attention modules to maintain the performance. Eventually, our dual-conditioned latent diffusion approach achieves better performance on most objective and subjective metrics than baselines on the AudioCaps test set. In the near future, we will extend this work to audio, speech, and music generation.

References

[1] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[4] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[5] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning*, pages 21450–21474. PMLR, 2023.

[6] Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*, 2024.

[7] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.

[8] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

[9] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[10] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

[11] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.

[12] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023.

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

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

[16] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

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

[18] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[20] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7441–7451, 2023.

[21] Weihao Weng and Xin Zhu. Inet: convolutional networks for biomedical image segmentation. *Ieee Access*, 9:16591–16603, 2021.

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

[23] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

[24] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. pages 2350–2354, 09 2019.

[25] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[26] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2019.