



# InfiniteAudio: Infinite-Length Audio Generation with Consistency

*Chaeyoung Jung, Hojoon Ki, Ji-Hoon Kim, Junmo Kim, Joon Son Chung*

Korea Advanced Institute of Science and Technology, South Korea

{codud9914, joon4366, jh.kim, junmo.kim, joonson}@kaist.ac.kr

## Abstract

This paper presents InfiniteAudio, a simple yet effective strategy for generating infinite-length audio using diffusion-based text-to-audio methods. Current approaches face memory constraints because the output size increases with input length, making long duration generation challenging. A common workaround is to concatenate short audio segments, but this often leads to inconsistencies due to the lack of shared temporal context. To address this, InfiniteAudio integrates seamlessly into existing pipelines without additional training. It introduces two key techniques: FIFO sampling, a first-in, first-out inference strategy with fixed-size inputs, and curved denoising, which selectively prioritizes key diffusion steps for efficiency. Experiments show that InfiniteAudio achieves comparable or superior performance across all metrics. Audio samples are available on our project page<sup>1</sup>.

**Index Terms:** text-to-audio generation, long generation, diffusion models

## 1. Introduction

Diffusion models [1, 2] have gained significant attention for their ability to generate high-quality and diverse outputs, achieving state-of-the-art performance across various domains, including image [3, 4, 5, 6], video [7, 8, 9, 10, 11, 12, 13], and audio [14, 15, 16, 17, 18, 19, 20]. However, their high computational cost limits their practicality in many applications. To address this, the Latent Diffusion Model (LDM) [4] was introduced, utilizing a compressed latent space to improve efficiency. This approach significantly reduces computational overhead while preserving high-fidelity outputs, making it more practical for image and video generation [21, 22, 23].

Beyond image and video generation, LDMs have also become a cornerstone in audio synthesis, particularly in text-to-audio (TTA) generation [24, 25, 26, 27, 28, 29, 30, 31, 32, 33], which produces realistic audio from textual prompts. By leveraging LDM's efficiency and generative power, TTA models have advanced significantly, incorporating Contrastive Language-Audio Pretraining (CLAP) [34] to enhance alignment between textual descriptions and generated audio [24, 25, 30]. Additionally, large language models (LLMs) have been integrated into TTA frameworks [27, 28], improving text embeddings and enabling more precise interpretation of complex prompts. This advancement allows TTA models to generate audio that more accurately reflects the intended context.

Despite significant advancements in TTA generation, existing models based on diffusion approaches face substantial challenges in generating long-duration audio [24, 26, 31, 35]. The

core issue arises from the inherent design of diffusion models, which require the input and output dimensions to remain identical throughout the process. As a result, generating longer audio necessitates a proportional increase in input size, leading to memory constraints. A common workaround involves concatenating short audio clips produced by existing TTA models to create longer sequences. However, this approach often suffers from temporal inconsistencies between segments, resulting in unnatural and discontinuous audio streams. The lack of shared temporal context across clips makes it difficult to maintain coherence and smooth transitions, further limiting the practical application of these models for long-form audio generation.

To address these limitations, we propose InfiniteAudio, a novel inference technique designed to generate long and temporally consistent audio. InfiniteAudio overcomes the memory constraints of diffusion models by employing a first-in-first-out (FIFO) mechanism with a fixed input size. This approach incrementally adds noise to parts of the existing model's predictions, as illustrated in Fig. 1. Unlike traditional diffusion models that begin with a uniform noise prior, InfiniteAudio uses priors with varying noise levels. During inference, the fully denoised portion of the input at the start is discarded, and new latent noise is appended at the end. This progressive replacement of input data enables InfiniteAudio to generate arbitrarily long audio sequences while maintaining temporal consistency and a constant memory footprint. By rethinking the inference process, InfiniteAudio eliminates the reliance on concatenating short audio clips and enables the direct generation of long-duration audio. This approach not only resolves the temporal inconsistencies of existing methods but also offers a solution for generating coherent, seamless audio over extended durations.

Although FIFO-Diffusion [36], a method developed for text-to-video (TTV) generation, employs a FIFO generation mechanism, it still utilizes all diffusion sampling steps during inference. In contrast, as illustrated in Fig. 1, our approach introduces curved denoising, a novel technique that selectively prioritizes the most critical diffusion steps identified through self-attention maps, rather than uniformly applying all steps. This targeted strategy preserves high-quality generation while substantially reducing the number of sampling steps, leading to a more efficient inference process.

Our contributions are summarized as follows. First, we first propose InfiniteAudio, a method capable of generating long-duration audio sequences without requiring additional training. It effectively addresses the memory limitations inherent in existing diffusion-based TTA models. Second, we introduce curved denoising, a selective sampling technique that focuses on critical diffusion steps instead of utilizing all steps, as in FIFO-Diffusion [36], resulting in improved sampling efficiency. Lastly, our method can be seamlessly integrated into existing

<sup>1</sup><https://mm.kaist.ac.kr/projects/InfiniteAudio/>

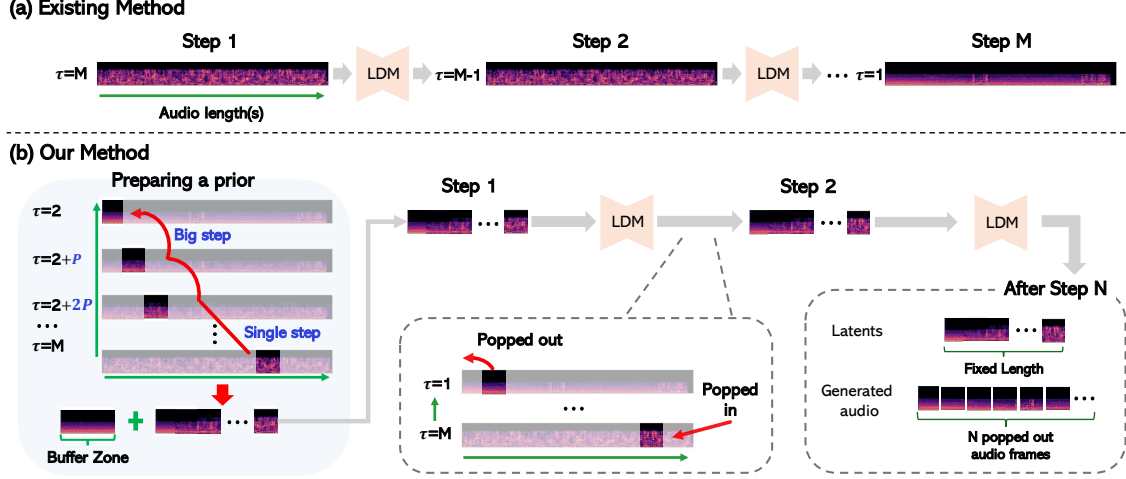


Figure 1: Comparison of Existing Methods and InfiniteAudio. Traditional methods apply uniform diffusion timesteps across all input latents, whereas InfiniteAudio dynamically selects timesteps based on their importance. This adaptive approach enables the generation of theoretically infinite audio while maintaining a fixed input size, ensuring both efficiency and high-quality synthesis.

text-to-audio generation baselines.

## 2. Method

### 2.1. Preliminaries

We provide a comprehensive overview of existing TTA generation models, which synthesize realistic audio from text prompts  $y$  by representing audio as a 2D mel-spectrogram, capturing both time and frequency dimensions.

Most TTA models share a common architecture, consisting of audio  $f_{audio}(\cdot)$  and text encoders  $f_{text}(\cdot)$ , a LDM, an audio decoder, and a vocoder. The encoders map text and audio inputs into a latent space, where the LDM is trained to iteratively refine a perturbed latent representation  $\mathbf{z}_\tau$ . Here,  $\tau \sim \mathcal{U}([1, \dots, M])$  represents the diffusion timestep, controlling the level of noise at each step. The LDM progressively denoises  $\mathbf{z}_\tau$  from a noisy state back to a clean latent representation  $\mathbf{z}_1$ . Once  $\mathbf{z}_1$  is obtained, the audio decoder reconstructs the mel-spectrogram  $a \in \mathbb{R}^{T \times F}$  from  $\mathbf{z}_1 \in \mathbb{R}^{C \times \frac{T}{r} \times \frac{F}{r}}$ , where  $T$  and  $F$  denote the time and frequency dimensions,  $C$  is the number of channels, and  $r$  is the compression factor. Finally, the vocoder converts the reconstructed mel-spectrogram into a waveform, producing the final audio output.

To train the model, Gaussian noise is gradually added to the latent representation over multiple timesteps. The model then learns to iteratively remove the noise to reconstruct the original clean representation. Given a random noise sample  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , where  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  denotes a standard normal distribution and a text condition  $\mathbf{c} = f_{text}(y)$  obtained from the text encoder, the model is trained to minimize the following denoising loss function:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, \epsilon, \tau} [|\epsilon - \epsilon_\theta(\mathbf{z}_\tau, \tau, \mathbf{c})|_2^2], \quad (1)$$

where  $\epsilon_\theta$  represents the model’s predicted noise at timestep  $\tau$ .

### 2.2. InfiniteAudio

In this section, we introduce InfiniteAudio, a technique for generating long-duration audio while maintaining a fixed memory footprint. Our approach leverages pre-trained diffusion-based TTA models and operates without requiring additional training.

We focus on two key models, AudioLDM and VoiceLDM, and demonstrate how our inference strategy effectively mitigates their inherent memory constraints.

#### 2.2.1. FIFO sampling

Generating long audio sequences with diffusion models is challenging due to their high memory requirements. Recently, FIFO-Diffusion [36] has addressed this issue in video generation by utilizing a fixed-size input, where each frame is assigned a different diffusion timestep. This allows the model to apply multiple diffusion steps simultaneously across the input frames. This approach enables the generation of theoretically infinite video by applying a FIFO sampling strategy.

We adapt this concept to audio generation by initiating the diffusion process with a fixed-length audio segment, where each segment is treated similarly to video frames. The input latent  $\mathbf{z}_\tau \in \mathbb{R}^{C \times \frac{T}{r} \times \frac{F}{r}}$  is treated as  $\frac{T}{r}$  audio frames, analogous to video frames. Here, each compressed mel-spectrogram frame corresponds to  $\mathbf{z}_1^i \in \mathbb{R}^{C \times 1 \times \frac{F}{r}}$ , where  $i \in [1, \frac{T}{r}]$ . This structure enables us to apply multiple diffusion steps across the audio segment in a similar manner to video generation.

For infinite audio generation, noise is progressively added to the input audio frames over time, except for the initial frames, which act as a “buffer zone” and are not perturbed by noise. Since no additional training occurs in InfiniteAudio, using different diffusion timesteps during inference can introduce a performance gap, as shown in [36]. The buffer zone helps mitigate this by ensuring that the same timesteps used during training are applied to the initial frames, reducing inconsistencies.

Beyond the buffer zone, the earlier frames are nearly fully predicted and the later frames are treated as Gaussian noise. At the inference stage, the input consists of the buffer frames and frames with increasing noise levels. As represented in Fig. 1 (b), after each inference step, the first frame following the buffer zone reaches diffusion timestep  $\tau = 1$  and is then removed. To maintain the same input size, we insert a new noisy frame at the last position. By iteratively repeating this process, we generate  $N$  frames in  $N$  inference steps, enabling seamless and consistent infinite audio generation through continuous synthesis.

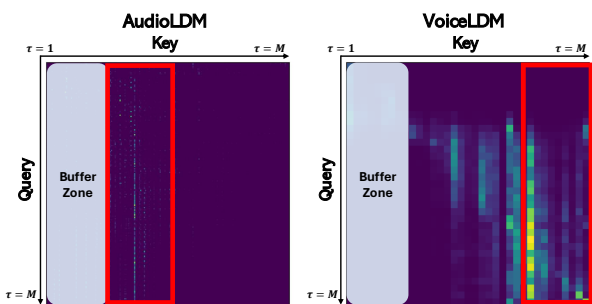


Figure 2: Attention maps indicating the importance of timesteps in input sequences. In AudioLDM, the query primarily attends to the initial portions of the key. In contrast, VoiceLDM exhibits a stronger correlation with the later key segments to its query, highlighting a distinct attention distribution pattern.

### 2.2.2. Curved Denoising with Reduced Sampling Steps

To address memory limitations, InfiniteAudio maintains a constant input size during inference, independent of the output length. However, using the full set of diffusion timesteps still requires long input sequences, as more audio segments must be generated by existing TTA models. To overcome this, InfiniteAudio prioritizes critical diffusion step regions while reducing emphasis on less important ones. By leveraging deterministic denoising [37], existing models can perform inference efficiently, skipping unnecessary steps while maintaining high-quality output. Similarly, we eliminate non-essential steps while preserving sample fidelity, guided by self-attention maps that highlight key regions for generation.

We partition the diffusion sampling steps into three distinct regions based on the original timestep distribution: initial, middle, and final. The initial region corresponds to the early stages of sampling, where diffusion timesteps are close to  $\tau$ , while the final region represents the later stages, where timesteps approach 1. To identify critical sampling regions during inference for AudioLDM and VoiceLDM, we analyze self-attention map scores within the InfiniteAudio framework. In the self-attention mechanism, attention scores quantify the relevance between a query and a key vector, determining the influence of one element on another within the sequence. By examining self-attention maps in U-Net decoder modules throughout the diffusion process, we can pinpoint key frames that exert the most influence on query frames in each model.

As shown in Fig. 2, model behavior varies significantly based on configuration. In AudioLDM, query sequences are primarily influenced by initial key sequences, corresponding to earlier frames with diffusion timesteps close to 1. In contrast, VoiceLDM exhibits greater sensitivity to later key sequences, which represent noisier inputs with timesteps approaching  $M$ . Additionally, some initial frames fall within a transitional buffer zone, so our analysis focuses on regions beyond this buffer for greater clarity. Based on these observations, we compute the average attention scores across the initial, middle, and final regions, as illustrated in Fig. 2. We then involve timesteps to regions with higher average attention scores, while skipping less critical regions using a factor of  $P$ . This adaptive timestep allocation effectively reduces both the number of inference steps and input size to about 3 seconds. Our curved denoising method preserves output quality while requiring fewer computations, enhancing overall efficiency.

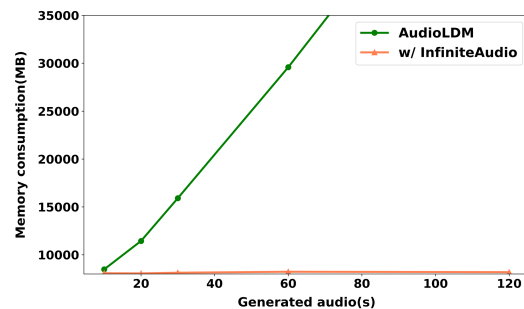


Figure 3: Memory consumption comparison between AudioLDM [24] and our method

## 3. Experiment

### 3.1. Experimental Settings

#### 3.1.1. Datasets and Baselines.

To evaluate our method on TTA generation, we utilize 500 audio-text pairs from the 975 test files in the Audiocaps dataset [38], which is commonly used for assessing TTA models. Given that our method relies heavily on the performance of existing baselines, we exclude the bottom 20 percent of audio-text pairs with the lowest CLAP scores, as predicted by AudioLDM and VoiceLDM. From the remaining pairs, 500 are randomly selected for evaluation. For comparison, we assess the performance of InfiniteAudio against two publicly available TTA models: AudioLDM<sup>2</sup> and VoiceLDM<sup>3</sup>.

#### 3.1.2. Evaluation Metrics.

We evaluate audio quality and text-audio alignment using standard quantitative metrics, including Fréchet Distance (FD), Kullback-Leibler (KL) divergence, and the CLAP score [24, 26, 39]. Fréchet Distance and Kullback-Leibler divergence quantify how closely the generated audio matches the ground truth, where lower values indicate better performance. The CLAP score measures the relevance of the generated audio to the text prompt, with higher values being preferable. For subjective evaluation, we assess overall quality (OVL) and relevance to the input text (REL). Both metrics were rated on a scale of 1 to 5 by 20 domain experts using 30 speech samples.

### 3.2. Quantitative Results

#### 3.2.1. Memory Consumptions.

We compare the memory consumption of AudioLDM with our method. AudioLDM’s memory usage grows with the length of the generated audio, while our method maintains constant memory usage regardless of audio length, as shown in Fig. 3.

#### 3.2.2. Evaluation of Curved Denoising.

We evaluate the effectiveness of our curved denoising strategy in Tab. 1, comparing 10-second audio samples generated using different sampling methods. Despite requiring no additional training, our method not only matches the performance of existing models but also achieves higher scores. By incorporating self-attention relevance, our approach outperforms methods that use equally spaced timesteps, such as FIFO-Diffusion [36], as well as other strategies with the same number of steps.

<sup>2</sup><https://github.com/haoheliu/AudioLDM>

<sup>3</sup><https://github.com/glory20h/VoiceLDM>

Table 1: *Quantitative evaluation of TTA. Our method demonstrates performance comparable to both models, even surpassing the original inference results. Additionally, the use of equally spaced timesteps, as suggested in [36], is considered.*

Method	CLAP $\uparrow$	FD $\downarrow$	KL $\downarrow$	OVL $\uparrow$	REL $\uparrow$
Ground Truth	0.5276	NA	NA	4.11 $\pm$ 0.22	4.03 $\pm$ 0.25
AudioLDM [24]	<b>0.4908</b>	44.6689	2.0805	<b>3.03<math>\pm</math>0.23</b>	<b>3.06<math>\pm</math>0.21</b>
w/ Equally spaced timesteps [36]	0.3832	54.7479	2.4013	2.19 $\pm$ 0.21	2.33 $\pm$ 0.23
w/ Middle focused timesteps	0.3979	56.7792	2.6077	2.06 $\pm$ 0.19	2.18 $\pm$ 0.20
<b>w/ Last focused timesteps (Ours)</b>	<b>0.4559</b>	<b>43.3788</b>	<b>1.9650</b>	<b>2.63<math>\pm</math>0.18</b>	<b>2.80<math>\pm</math>0.21</b>
w/ Initial focused timesteps	0.3110	67.0704	2.9838	2.13 $\pm$ 0.19	2.07 $\pm$ 0.20
VoiceLDM [26]	<b>0.4199</b>	<b>51.4019</b>	<b>2.2749</b>	<b>2.53<math>\pm</math>0.24</b>	2.41 $\pm$ 0.21
w/ Equally spaced timesteps [36]	0.3729	59.1521	2.4477	2.20 $\pm$ 0.21	2.33 $\pm$ 0.22
w/ Middle focused timesteps	0.3779	56.7321	2.4622	2.10 $\pm$ 0.20	2.41 $\pm$ 0.22
w/ Last focused timesteps	0.3542	64.8813	2.6227	<b>2.38<math>\pm</math>0.23</b>	2.24 $\pm$ 0.21
<b>w/ Initial focused timesteps (Ours)</b>	<b>0.4107</b>	<b>51.5047</b>	<b>2.3498</b>	<b>2.38<math>\pm</math>0.23</b>	<b>2.48<math>\pm</math>0.21</b>

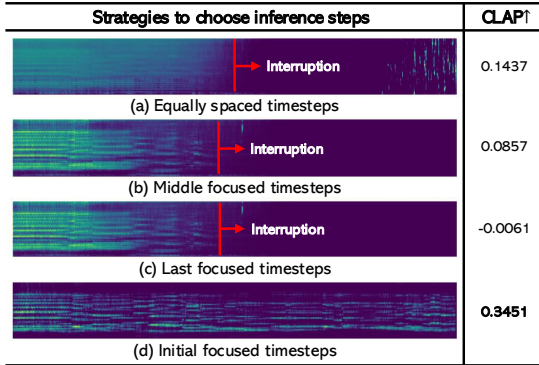


Figure 4: *Analysis of different diffusion sampling strategies for VoiceLDM [26].*

Table 2: *Comparison of sampling steps in VoiceLDM [26]. InfiniteAudio achieves superior results for 10 second audio generation while requiring fewer than 150 sampling steps, demonstrating enhanced efficiency.*

Sampling steps	CLAP $\uparrow$	FD $\downarrow$	KL $\downarrow$
w/ 200 equally spaced steps	0.3923	53.0555	<b>2.3334</b>
w/ 250 equally spaced steps	0.3941	<b>50.5447</b>	2.3937
<b>InfiniteAudio</b>	<b>0.4107</b>	<b>51.5047</b>	<b>2.3498</b>

### 3.3. Qualitative Results

#### 3.3.1. Sampling Strategies.

As shown in Fig. 4, unlike other strategies that exhibit interruptions in the generated audio, as seen in the spectrograms, our method ensures seamless audio generation. This is supported by both the spectrogram analysis and the improved CLAP score.

#### 3.3.2. Long Generation Quality.

Fig. 5 presents the mel spectrograms of audio generated by AudioLDM [24] at different lengths using InfiniteAudio, compared to audio produced via the concatenation method. The results demonstrate that InfiniteAudio generates high-quality, long-duration audio while preserving stable CLAP scores across varying lengths. Moreover, it maintains seamless consistency and natural coherence throughout the entire audio, whereas the concatenation method results in repetitive patterns and temporal inconsistencies, as highlighted by the red line.

### 3.4. Analysis on Sampling Steps and Audio Length

InfiniteAudio is designed to optimize sampling efficiency by minimizing the number of sampling steps while maintaining high audio quality. Unlike methods that extend sampling to

Text prompts	Length	Method	Spectrograms for generated Audio	CLAP $\uparrow$
Trumpet	20s	Ours		0.3888
		Ours		<b>0.4964</b>
	60s	Concat		<b>0.3927</b>
A capella	20s	Ours		<b>0.2995</b>
		Ours		<b>0.2693</b>
	60s	Concat		<b>0.2958</b>

Figure 5: *Comparison of audio generated by AudioLDM [24] using InfiniteAudio and the concatenation method, demonstrating InfiniteAudio’s superior long-duration generation.*

Table 3: *Comparison of generated audio lengths between a fixed 10 second and variable-length in AudioLDM [24].*

Generated audio length	CLAP $\uparrow$	FD $\downarrow$	KL $\downarrow$
Fix	0.3207	<b>43.3788</b>	1.9650
Various	<b>0.3259</b>	44.3701	<b>1.9044</b>

200 or 250 steps with equally spaced timesteps, our approach achieves consistently solid performance across all metrics while using fewer than 150 steps, as shown in Tab. 2.

InfiniteAudio also excels in generating longer audio sequences without sacrificing quality. As shown in Tab. 3, it delivers results comparable to the fixed 10-second generation across varying lengths, ranging from 10 to 20 seconds. Notably, the CLAP score for this experiment is calculated using a different checkpoint than the one used in other tables, as evaluating varying audio lengths requires a distinct CLAP model<sup>4</sup>.

## 4. Conclusion

We introduce InfiniteAudio, a novel inference method for generating infinitely long, consistent audio using pretrained text-to-audio models. By maintaining a fixed memory footprint, InfiniteAudio overcomes memory constraints in existing models and integrates seamlessly with diffusion-based TTA approaches. Despite relying solely on inference techniques, it achieves superior performance, opening new possibilities for continuous and coherent long-form audio generation.

## 5. Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00212845, Multimodal Speech Processing for Human-Computer Interaction).

<sup>4</sup><https://github.com/LAION-AI/CLAP>

## 6. References

- [1] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *NeurIPS*, 2020.
- [2] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *Proc. ICLR*, 2021.
- [3] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *NeurIPS*, 2021.
- [4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. CVPR*, 2022.
- [5] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” in *NeurIPS*, 2022.
- [6] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, “Glide: Towards photorealistic image generation and editing with text-guided diffusion models,” *arXiv:2112.10741*, 2021.
- [7] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, “Video diffusion models,” *NeurIPS*, 2022.
- [8] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, “Make-a-video: Text-to-video generation without text-video data,” *arXiv:2209.14792*, 2022.
- [9] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, “Modelscope text-to-video technical report,” *arXiv:2308.06571*, 2023.
- [10] R. Yang, P. Srivastava, and S. Mandt, “Diffusion probabilistic modeling for video generation,” *Entropy*, vol. 25, p. 1469, 2023.
- [11] Y. Wang, X. Chen, X. Ma, S. Zhou, Z. Huang, Y. Wang, C. Yang, Y. He, J. Yu, P. Yang *et al.*, “Lavie: High-quality video generation with cascaded latent diffusion models,” *International Journal of Computer Vision*, vol. 133, pp. 3059–3078, 2024.
- [12] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, G. Liu, A. Raj *et al.*, “Lumiere: A space-time diffusion model for video generation,” in *SIGGRAPH Asia 2024 Conference Papers*, 2024.
- [13] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang *et al.*, “Videocrafter1: Open diffusion models for high-quality video generation,” *arXiv:2310.19512*, 2023.
- [14] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, “Grad-tts: A diffusion probabilistic model for text-to-speech,” in *Proc. ICML*, 2021.
- [15] H. Kim, S. Kim, and S. Yoon, “Guided-tts: A diffusion model for text-to-speech via classifier guidance,” in *Proc. ICML*, 2022.
- [16] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, “Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models,” *NeurIPS*, 2024.
- [17] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, “Diff-tts: A denoising diffusion model for text-to-speech,” in *Proc. Interspeech*, 2021.
- [18] Z. Liu, Y. Guo, and K. Yu, “Diffvoice: Text-to-speech with latent diffusion,” in *Proc. ICASSP*, 2023.
- [19] S. Lee, C. Jung, Y. Jang, J. Kim, and J. S. Chung, “Seeing through the conversation: Audio-visual speech separation based on diffusion model,” in *Proc. ICASSP*, 2024.
- [20] C. Jung, S. Lee, J.-H. Kim, and J. S. Chung, “Flowavse: Efficient audio-visual speech enhancement with conditional flow matching,” in *Proc. Interspeech*, 2024.
- [21] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, “Align your latents: High-resolution video synthesis with latent diffusion models,” in *Proc. CVPR*, 2023.
- [22] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “Sdxl: Improving latent diffusion models for high-resolution image synthesis,” in *Proc. ICLR*, 2025.
- [23] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, “Mag-icvideo: Efficient video generation with latent diffusion models,” *arXiv:2211.11018*, 2022.
- [24] 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.
- [25] 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.
- [26] Y. Lee, I. Yeon, J. Nam, and J. S. Chung, “Voiceldm: Text-to-speech with environmental context,” in *Proc. ICASSP*, 2024.
- [27] 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 Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.
- [28] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, “Text-to-audio generation using instruction-tuned llm and latent diffusion model,” in *Proc. ACM MM*, 2023.
- [29] 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 Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023.
- [30] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, “Retrieval-augmented text-to-audio generation,” in *Proc. ICASSP*, 2024.
- [31] 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.
- [32] H. Liu, R. Huang, Y. Liu, H. Cao, J. Wang, X. Cheng, S. Zheng, and Z. Zhao, “Audioldm: Text-to-audio generation with latent consistency models,” in *Proc. ACM MM*, 2024.
- [33] J. Jung, J. Ahn, C. Jung, T. D. Nguyen, Y. Jang, and J. S. Chung, “Voicedit: Dual-condition diffusion transformer for environment-aware speech synthesis,” in *Proc. ICASSP*, 2025.
- [34] 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*, 2023.
- [35] Z. Guo, J. Mao, R. Tao, L. Yan, K. Ouchi, H. Liu, and X. Wang, “Audio generation with multiple conditional diffusion model,” in *Proc. AAAI*, 2024.
- [36] J. Kim, J. Kang, J. Choi, and B. Han, “Fifo-diffusion: Generating infinite videos from text without training,” in *NeurIPS*, 2024.
- [37] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *Proc. ICLR*, 2021.
- [38] C. D. Kim, B. Kim, H. Lee, and G. Kim, “AudioCaps: Generating captions for audios in the wild,” in *NAACL-HLT*, 2019.
- [39] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan *et al.*, “Audiobox: Unified audio generation with natural language prompts,” *arXiv:2312.15821*, 2023.