Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

EzAudio: Enhancing Text-to-Audio Generation with Efficient Diffusion Transformer

Jiarui Hai[*1], *Yong Xu*[2], *Hao Zhang*[2], *Chenxing Li*[2], *Helin Wang*[1], *Mounya Elhilali*[1], *Dong Yu*[2]

[1]Department of Electrical and Computer Engineering, Johns Hopkins University, MD, USA
[2]AI Lab, Tencent Americas, WA, USA

jhai2@jhu.edu, lucayongxu@global.tencent.com

Abstract

We introduce EzAudio, a text-to-audio (T2A) generation framework designed to produce high-quality, natural-sounding sound effects. Core designs include: (1) We propose EzAudio-DiT, an optimized Diffusion Transformer (DiT) designed for audio latent representations, improving convergence speed, as well as parameter and memory efficiency. (2) We apply a classifier-free guidance (CFG) rescaling technique to mitigate fidelity loss at higher CFG scores and enhancing prompt adherence without compromising audio quality. (3) We propose a synthetic caption generation strategy leveraging recent advances in audio understanding and LLMs to enhance T2A pretraining. We show that EzAudio, with its computationally efficient architecture and fast convergence, is a competitive open-source model that excels in both objective and subjective evaluations by delivering highly realistic listening experiences.

Index Terms: text-to-audio generation, diffusion model, diffusion transformer

1. Introduction

The rapid advancement of diffusion-based generative models has transformed content creation, particularly in image synthesis [1]. Inspired by this, early text-to-audio (T2A) methods used spectrogram-based representations, evolving into a powerful approach for high-quality sound generation [2, 3, 4]. Recent work [5, 6, 7, 8] has improved T2A quality by adopting one-dimensional (1D) latent audio representations. The Diffusion Transformer (DiT) [9], leveraging Adaptive LayerNorm (AdaLN) for diffusion step fusion, has shown strong performance in visual generation and, more recently, in sound generation [5]. Despite these advances, recent T2A pipelines still have room for improvement: (1) DiT in audio generation requires substantial memory and training costs and could benefit from further optimization for T2A tasks and latent audio representations. (2) We find the T2A model using waveform latents exhibit noticeable fidelity loss at high classifier-free guidance (CFG) [10] scores. While higher guidance improves prompt coherence, it can distort the waveform amplitude distribution, subsequently affecting frequency features and introducing artifacts, leading to degradation in generation quality.

Beyond model design, pretraining plays a crucial role in achieving high-quality T2A due to the scarcity of human-labeled data. Strategies [7, 11] have been proposed to leverage unlabeled data for representation learning and improving generation quality. However, text-to-audio mapping pretraining strategies still face challenges. Using CLAP embeddings [3] derived from unlabeled audio data and switching to text-derived CLAP embeddings in downstream tasks can limit performance due to mismatches between audio and text representations. Tagging-based pseudo captions [2, 12, 7] directly incorporate text conditions during pretraining but lack sequential information about sound events, limiting the model's ability to process fine-grained prompts in downstream tasks. Synthetic audio data [4, 6] offers precise descriptions and timing alignment but is difficult to prepare and may introduce artifacts or unnatural characteristics due to discrepancies with real audio.

To address these challenges, we propose following core designs: (1) **EzAudio-DiT**, an optimized DiT architecture for efficient, high-quality T2A. It features a novel AdaLN variant that reduces parameters and memory consumption without compromising performance, along with long-skip connections to accelerate convergence. (2) We enhance CFG sampling by adopting **CFG rescaling** [13], originally developed to prevent overexposure in image generation. We demonstrate that when applied to waveform latents, it mitigates fidelity degradation at high CFG scores while preserving strong prompt adherence, eliminating the need for meticulous CFG tuning. (3) We leverage recent advances in audio understanding and LLMs to generate high-quality **synthetic caption** data for efficient T2A pretraining. Specifically, we prepare the following sources of synthetic caption data: (a) generating captions using audio captioning and audio-language models, which have demonstrated the ability to interpret complex auditory scenes [14, 15, 16], and (b) enriching strong sound event labels [17] with LLMs to generate captions that provide detailed sequential information about sound events.

As a result of these designs, **EzAudio**[1] achieves fast convergence with reduced parameters and memory usage and is able to generate highly realistic audio. It stands out as a competitive open-source model in both objective and subjective evaluations. We hope our model and pipeline empower researchers and startups to develop T2A models more effectively and efficiently.

2. Method

EzAudio builds on recent advances in diffusion-based audio and music generation [2, 5]. As shown in Figure 1, it comprises three components: (1) a FLAN-T5-based text encoder [18] for processing audio descriptions, (2) a latent diffusion model for generating audio latents, and (3) a waveform VAE [5] for reconstructing waveforms from audio latents.

The following sections detail EzAudio's core designs: Section 2.1 introduces the proposed EzAudio-DiT for diffusion modeling, Section 2.2 describes CFG rescaling for sampling, and Section 2.3 covers data labeling and training strategy.

[*]Work done during J. Hai's internship at Tencent AI Lab, USA.

[1]Code and demo: https://haidog-yaqub.github.io/EzAudio-Page/

10.21437/Interspeech.2025-1137

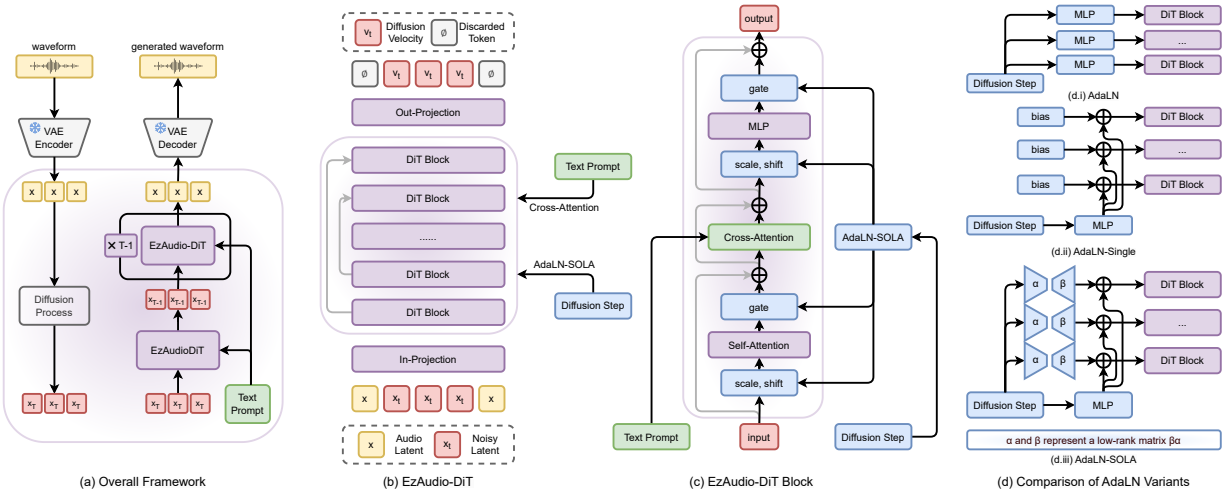


Figure 1: *The framework of EzAudio and the architectural details of EzAudio-DiT.*

2.1. Efficient EzAudio-DiT

Stable Audio [5] has successfully used DiT [9] for text-to-music generation. However, we find opportunities to optimize DiT's efficiency and convergence speed in audio generation. To achieve this, we propose two key modifications that enhance parameter and memory efficiency while accelerating convergence, without compromising training stability. These designs include:

AdaLN-SOLA: The AdaLN layers in DiT are crucial for managing both image class conditions and diffusion steps but account for a significant portion of the model's parameters. However, in T2A, where cross-attention processes text conditions, AdaLN can be simplified. AdaLN-Single [11] addresses this by sharing a single AdaLN module across all DiT blocks but degrades performance and even destabilizes training. To address this, we propose **AdaLN-SOLA** (**AdaLN-S**ingle **O**rchestrated by **L**ow-rank **A**djustment), inspired by low-rank adaption methods[19]. As shown in Figure 1 (d), AdaLN-SOLA retains a shared AdaLN module but incorporates block-specific low-rank matrices that dynamically adjust it based on diffusion steps. This approach reduces parameters and memory usage while preserving performance and stability.

Long-skip Connection: Earlier diffusion models use long-skip connections to propagate low-level features and diffusion steps into deeper layers for effective modeling. Recent DiT architectures [11, 5] remove these connections, relying on transformer residuals for feature propagation and assuming AdaLN can handle the diffusion steps. However, we find that removing skip connections slows convergence and degrades performance, particularly for waveform latent embeddings with 128 channels—far more than typical image representations—making them difficult to process solely through residual connections. To address this, we integrate long-skip connections into DiT, inspired by U-ViT designs [20, 21, 7], allowing low-level features to directly reach deeper transformer blocks, as shown in Figure 1 (b).

2.2. CFG Rescaling

The CFG [10] is used to direct the diffusion sampling. It modifies the output v only during the reverse process according to:

$$v_{cfg} = v_{neg} + w(v_{pos} - v_{neg}), \tag{1}$$

where w is the guidance scale, and v_{pos} and v_{neg} represent model outputs under positive and negative prompts, with v_{cfg} being the adjusted velocity. By default, the negative prompt is set to empty, corresponding to the unconditional case.

A higher guidance scale enhances prompt alignment but can disrupt the waveform's amplitude distribution, affecting frequency characteristics and ultimately degrading generation quality. The CFG rescaling technique [13] is used to adjust the magnitude of v_{cfg} while preserving its direction when a large w is employed.

$$v_{re} = v_{cfg} \cdot \mathrm{std}(v_{pos}) \cdot \mathrm{std}(v_{cfg})^{-1}, \tag{2}$$

$$v'_{cfg} = \phi \cdot v_{re} + (1 - \phi) \cdot v_{cfg}, \tag{3}$$

where ϕ is the rescaling factor, with v'_{cfg} denoting the refined CFG velocity for diffusion sampling.

2.3. Training Strategy

Synthetic Caption Data Generation: We utilize multiple approaches to generate synthetic caption data, enhancing caption diversity and richness: **(1) Auto-ACD** [22] utilizes audio and video captioning models to generate initial captions, which are then refined by a language model into natural audio descriptions for AudioSet and VGGSound. **(2) AS-Qwen-Caps** uses Qwen-Audio[2] [14], one of the leading audio-language models, to describe audio from AudioSet and VGGSound; **(3) AS-SL-GPT4-Caps** uses OpenAI's GPT-4o-mini API[3] to prepare descriptions that emphasize sequential information based on temporal annotations from the strongly labeled subset of AudioSet [17]. To ensure the quality and accuracy of captions, we use a CapFilt-like [23, 16] filtering method, leveraging a pre-trained CLAP model [24] to discard audio-caption pairs with similarity scores below a set threshold.

Multi-Stage Training: We adopt a three-stage training approach [11, 7] to leverage unlabeled audio data and enhance generation quality. **(1) Masked Audio Modeling:** Following diffusion-based mask pretraining methods [21, 7, 25], the diffusion model is first trained to predict masked tokens from unmasked ones, without text conditioning. A random portion of

[2]We compare Qwen-Audio [14] and GAMA [15], selecting Qwen-Audio for its higher accuracy and fewer hallucinations on AudioCaps.

[3]https://platform.openai.com/docs/models/gpt-4

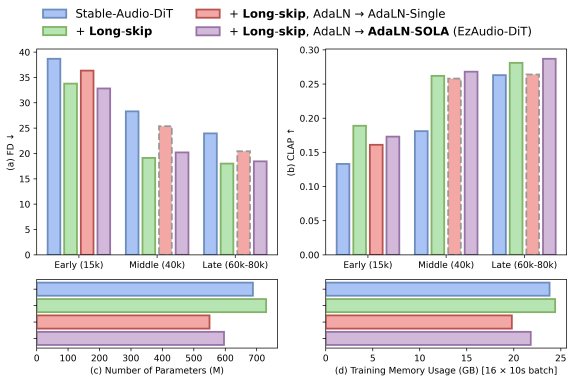


Figure 2: *Ablation of DiT Design. The gray dashed edges indicate results from training resumed after a crash.*

tokens—ranging from 25% to 100% with a minimum span of 0.2s—is masked, and the cross-attention module in transformer blocks is excluded during this stage. **(2) Text-Audio Alignment:** This stage integrates synthetic captions to facilitate text-audio alignment learning. Building on the masked modeling stage, we introduce a randomly initialized cross-attention module into each DiT block to process text conditions. To ensure a smooth training transition, we initialize the output projection layer of the cross-attention module to zero. Additionally, to encourage greater reliance on text input, we set a fixed 75% probability of fully masking all tokens during training. **(3) Supervised Fine-Tuning:** Finally, following Tango [2], we fine-tune the model on AudioCaps [26] to further enhance performance.

3. Experiments

3.1. Experimental Setups

We conduct experiments using a 24kHz sample rate for both the waveform VAE and the T2A model. The waveform latent operate at 50Hz and consists of 128 channels. We train the waveform VAE on AudioSet [27] for 1 million steps, enabling it to handle a wide variety types of sounds. For DiT models, DiT-L consists of 24 DiT blocks, each with 1024 channels, while DiT-XL has 28 DiT blocks, each with 1152 channels. The rank in AdaLN-SOLA is 32 for DiT-L and 36 for DiT-XL. The LDM employs velocity (v) prediction and Zero-SNR schedulers [13], both effective in diffusion-based image and audio generation [28, 29]. We use 50 sampling steps and a CFG score of 3 by default in the ablation studies presented in Sections 3.2 and 3.3.

Following previous T2A studies [3, 30, 2, 6], we evaluate our model using Frechet Distance (FD)[4], Kullback–Leibler (KL) divergence, and Inception Score (IS), with pre-trained PANNs [31] as the feature extractor. Additionally, we employ CLAP[5] [24] to assess the coherence between the generated audio and the text prompt. All audio samples are **resampled to 16kHz** during evaluation. The AudioCaps test set, comprising 900 audio clips with 882 currently available, is used for evaluation. Each clip has five captions, and we **randomly select one caption per clip**, following AudioLDM and Tango [2, 3].

[4]We exclude FAD due to reliability concerns [3, 16].

[5]https://huggingface.co/laion/larger_clap_general

Table 1: *Comparison of pretraining methods.*

Dataset	Mask Mod.	FD↓	KL↓	IS↑	CLAP↑
WavCaps	No	17.79	1.66	9.60	0.273
EzAudioCaps	No	16.60	1.67	10.04	0.288
EzAudioCaps	**Yes**	**15.46**	**1.44**	**10.11**	**0.294**

Table 2: *Ablation of CLAP filtering.*

Threshold	# Samples	FD↓	KL↓	IS↑	CLAP↑
0.35	0.58M	<u>16.17</u>	1.48	9.85	0.290
0.45	0.11M	16.27	**1.40**	**10.31**	**0.303**
0.40	0.27M	**15.46**	<u>1.44</u>	<u>10.11</u>	<u>0.294</u>

3.2. Ablation of DiT Architecture

We perform an ablation study on different DiT designs using the AudioCaps dataset, training for 80k steps with a batch size of 128 and a learning rate of 1e-4, following the DiT-L configuration in Section 3.1. Stable-Audio-DiT [5], which extends DiT [9] with cross-attention and Rotary Position Embedding (RoPE) [32], serves as our baseline. We investigate the effects of adding long-skip connections and replacing AdaLN with either AdaLN-Single or the proposed AdaLN-SOLA. We compare convergence across three training stages. Model performance improves steadily during the early and middle stages, with results reported at 15k and 40k steps. In the late stage, performance stabilizes with minor fluctuations, and the best scores between 60k and 80k steps are reported based on validation loss.

The key findings can be summarized as follows: (1) As shown in Figure 2 (a) and (b), long-skip connections significantly accelerate convergence and lead to better model performance; (2) Replacing AdaLN with AdaLN-Single leads to performance degradation and introduces numerical instability, causing training crashes, whereas AdaLN-SOLA maintains stability with minimal impact on performance; (3) Figure 2 (c) and (d) illustrates that long-skip connections slightly increase model parameters and memory usage, while AdaLN-SOLA substantially reduces both, resulting in a more lightweight model. In conclusion, EzAudio-DiT achieves faster convergence and greater efficiency than Stable-Audio-DiT.

3.3. Ablation of Training Strategy

We conduct an ablation study comparing our dataset with WavCaps [12], which enriches audio tags using ChatGPT but often lacks sequential information[6] and has been used for pretraining in Tango-Full [2]. Additionally, we evaluate the impact of mask modeling. For pretraining without mask modeling, we use a batch size of 128 and train for 150K steps at a 1e-4 learning rate with synthetic caption data, followed by 30K fine-tuning steps on AudioCaps at 1e-5. When incorporating mask modeling, we first train on AudioSet for 100K steps at 1e-4, then perform 50K steps on synthetic caption data at 5e-5, and conclude with 30K fine-tuning steps at 1e-5.

As shown in Table 1, our proposed dataset improves generation quality and strengthens text coherence. Also, mask modeling pretraining further enhances overall generation quality.

Additionally, we evaluate different thresholds for filtering synthetic captions. The threshold selection is based on the mean CLAP score of AudioCaps, which is around 0.30. We

[6]Comparison with our dataset available on the Demo page.

Table 3: *Comparison of EzAudio and T2A models with evaluation results on AudioCaps. † denotes trainable parameters.*

Method	Model	# Params.†	Pretrain Data	Text Encoder	FD↓	KL↓	IS↑	CLAP↑
Ground Truth	–	–	–	–	–	–	–	0.302
Tango [2]	2D U-Net	866M	Synthetic Caption	FLAN-T5	19.07	1.33	7.70	0.293
Tango-AF [16]	2D U-Net	866M	Synthetic Caption	FLAN-T5	21.84	1.32	9.20	0.269
AudioLDM[10] [3]	2D U-Net	739M	CLAP Embedding	CLAP	30.96	2.36	7.38	0.197
AudioLDM-2[10] [30]	2D U-Net	712M	Synthetic Caption	CLAP + FLAN-T5	25.03	1.75	8.13	0.236
Make-An-Audio [4]	2D U-Net	453M	Synthetic Audio	CLAP	18.77	1.71	8.80	0.244
Make-An-Audio-2[11] [6]	1D Transformer	937M	Synthetic Audio	CLAP + FLAN-T5	16.16	1.42	9.93	0.284
Gen-AU-Large [8]	1D Transformer	1.25B	Synthetic Caption	CLAP + FLAN-T5	17.21	1.40	**11.42**	0.270[9]
EzAudio-L	1D Transformer	596M	Synthetic Caption	FLAN-T5	15.59	1.38	11.35	**0.319**
EzAudio-XL	1D Transformer	874M	Synthetic Caption	FLAN-T5	**14.98**	**1.29**	11.38	0.314

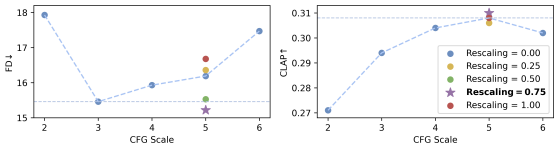


Figure 3: *Ablation of CFG scales and rescaling factors.*

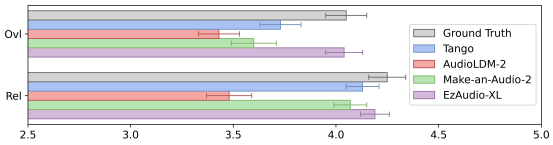


Figure 4: *Mean opinion scores with 95% confidence intervals.*

set higher thresholds than this number to prioritize quality. All models are trained with mask pretraining. As shown in Table 2, a lower threshold allows for more diverse but noisier data, negatively impacting all metrics, whereas a higher threshold improves most metrics but reduces FD and limits data diversity. We adopt a threshold of 0.40 for EzAudioCaps, as it provides the best balance between data diversity and model performance.

3.4. Ablation of CFG and CFG Rescaling

As shown in Figure 3, we evaluate CFG scores using the model trained in Section 3.3. Higher CFG values improve text-audio alignment but also increase FD, indicating a decline in audio quality. With CFG = 5 yielding the highest CLAP score and less FD degradation, we apply rescaling at this level. A rescaling factor around 0.50–0.75 helps maintain strong prompt alignment while mitigating the negative impact on audio quality.

3.5. Comparison with State-of-the-art

We compare EzAudio with recent open-source T2A models, introducing two variants: EzAudio-L and EzAudio-XL, which differ only in model size, corresponding to EzAudio-DiT in L and XL configurations, as described in Section 3.1. Both models are trained on the proposed dataset using mask modeling, as detailed in Section 3.3. We use a CFG score of 5 and a rescaling score of 0.75, as stated in Section 3, while increasing the sampling steps to 100.

The baseline models[7] are summarized in Table 3. To ensure a fair comparison, we use the official checkpoints[8] or provided samples[9] for each baseline. All models are evaluated using the same test method and dataset described in Section 3.1, following the recommended sampling configurations from their respective papers or repositories.

As shown in Table 3, 2D U-Net-based models[10] perform worse on FD and IS metrics, producing less realistic audio. Among them, Tango stands out with a strong CLAP score, indicating better coherence. More recent models[11] leveraging a 1D VAE and transformer architecture demonstrate notable improvements in FD and IS. In particular, Gen-AU-Large, benefiting from a larger model scale and extensive pre-training, further enhances audio quality, achieving the highest IS. EzAudio-L and EzAudio-XL match or surpass baselines across various metrics, highlighting their superior quality and prompt coherence, with EzAudio-XL holding a slight overall advantage.

We conduct a subjective experiment[12] to evaluate overall audio quality (OVL) and text relevance (REL) using a 5-point Mean Opinion Score (MOS) on 30 randomly selected prompts. 12 participants with backgrounds in music production or sound engineering take part. As shown in Figure 4, results align with objective findings: EzAudio-XL outperforms baselines in both relevance and quality. Make-An-Audio 2 receives a lower OVL score than its objective metrics suggest, likely due to artifacts from synthetic data. Notably, EzAudio-XL's OVL score approaches real recordings, demonstrating its ability to generate highly realistic audio.

4. Conclusion

In this paper, we introduce EzAudio, a framework that integrates a training- and computationally-efficient DiT architecture, an effective training pipeline leveraging synthetic caption data, and CFG rescaling to achieve precise and high-quality audio generation. In future work, we plan to incorporate techniques such as ControlNet, and DPO to further improve controllability and generation quality.

[7]Stable Audio focuses on music generation, so we exclude it from the final T2A comparison but compare its DiT in Section 3.2.

[8]For baselines with multiple versions, we use *tango-full-ft-ac*, *tango-af-ac-ft-ac*, *audioldm-l-full*, and *audioldm2-large*.

[9]Gen-AU releases samples with YouTube IDs but no captions. Since each ID can correspond to five different captions, linking a sample to its original prompt caption isn't always accurate, which may affect the precision of the CLAP score.

[10]The open-source AudioLDMs lack fine-tuning or exclusive training on AudioCaps, leading to differences from the paper's best results.

[11]Make-An-Audio 2 uses all five captions per clip to compute metrics, whereas Tango, AudioLDM, and our method use one randomly selected caption, leading to differences with its reported results.

[12]Due to cost constraints, we compare EzAudio-XL with Tango, AudioLDM-2, Make-An-Audio-2, and real samples from AudioCaps.

5. References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] 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.

[3] 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 *International Conference on Machine Learning*. PMLR, 2023, pp. 21 450–21 474.

[4] 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 *International Conference on Machine Learning*. PMLR, 2023, pp. 13 916–13 932.

[5] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Long-form music generation with latent diffusion," *arXiv preprint arXiv:2404.10301*, 2024.

[6] 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.

[7] 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 preprint arXiv:2312.15821*, 2023.

[8] M. Haji-Ali, W. Menapace, A. Siarohin, G. Balakrishnan, S. Tulyakov, and V. Ordonez, "Taming data and transformers for audio generation," *arXiv preprint arXiv:2406.19388*, 2024.

[9] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

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

[11] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Z. Wang, J. T. Kwok, P. Luo, H. Lu, and Z. Li, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.

[12] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[13] S. Lin, B. Liu, J. Li, and X. Yang, "Common diffusion noise schedules and sample steps are flawed," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 5404–5411.

[14] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[15] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities," *arXiv preprint arXiv:2406.11768*, 2024.

[16] Z. Kong, S.-g. Lee, D. Ghosal, N. Majumder, A. Mehrish, R. Valle, S. Poria, and B. Catanzaro, "Improving text-to-audio models with synthetic captions," *arXiv preprint arXiv:2406.15487*, 2024.

[17] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[18] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[20] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A vit backbone for diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 669–22 679.

[21] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *Advances in neural information processing systems*, vol. 36, 2024.

[22] L. Sun, X. Xu, M. Wu, and W. Xie, "Auto-ACD: A large-scale dataset for audio-language representation learning," in *ACM Multimedia 2024*, 2024.

[23] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[24] 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 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] S. Gao, P. Zhou, M.-M. Cheng, and S. Yan, "Masked diffusion transformer is a strong image synthesizer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 164–23 173.

[26] C. D. Kim, B. Kim, H. Lee, and G. 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)*, 2019, pp. 119–132.

[27] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[28] 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 *The Twelfth International Conference on Learning Representations*, 2024.

[29] J. Hai, H. Wang, D. Yang, K. Thakkar, N. Dehak, and M. Elhilali, "Dpm-tse: A diffusion probabilistic model for target sound extraction," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1196–1200.

[30] 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*, 2024.

[31] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[32] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.