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], and Dong Yu[2]
[1]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
[2]Tencent AI Lab, Bellevue, WA, USA
Email: jhai2@jhu.edu, lucayongxu@global.tencent.com, mounya@jhu.edu

***Abstract*—Latent diffusion models have shown promising results in text-to-audio (T2A) generation tasks, yet previous models have encountered difficulties in generation quality, computational cost, diffusion sampling, and data preparation. In this paper, we introduce EzAudio, a transformer-based T2A diffusion model, to handle these challenges. Our approach includes several key innovations: (1) We build the T2A model on the latent space of a 1D waveform Variational Autoencoder (VAE), avoiding the complexities of handling 2D spectrogram representations and using an additional neural vocoder. (2) We design an optimized diffusion transformer architecture specifically tailored for audio latent representations and diffusion modeling, which enhances convergence speed, training stability, and memory usage, making the training process easier and more efficient. (3) To tackle data scarcity, we adopt a data-efficient training strategy that leverages unlabeled data for learning acoustic dependencies, audio caption data annotated by audio-language models for text-to-audio alignment learning, and human-labeled data for fine-tuning. (4) We introduce a classifier-free guidance (CFG) rescaling method that simplifies EzAudio by achieving strong prompt alignment while preserving great audio quality when using larger CFG scores, eliminating the need to struggle with finding the optimal CFG score to balance this trade-off. EzAudio surpasses existing open-source models in both objective metrics and subjective evaluations, delivering realistic listening experiences while maintaining a streamlined model structure, low training costs, and an easy-to-follow training pipeline. Code, data, and pre-trained models are released at: https://haidog-yaqub.github.io/EzAudio-Page/.**

***Index Terms*—Diffusion transformers, text-to-audio generation**

I. Introduction

The rapid advancement of latent diffusion-based text-to-image (T2I) generative models, such as Stable Diffusion [1], has revolutionized the field of high-quality image synthesis. Building on the success of these methods, diffusion-based text-to-audio (T2A) generation has emerged as a promising area of research. Previous studies [2]–[6] adapted techniques from T2I by treating 2D mel spectrograms as images, utilizing 2D U-Nets to generate audio content.

Recently, the Diffusion Transformer (DiT) [7]–[9] has shown outstanding performance in image generation, surpassing traditional CNN-based U-Nets. However, when applied to T2A, particularly those involving 2D mel spectrograms, it faces challenges. A key issue is balancing computational cost and temporal resolution. Also, 2D-mel based methods could not be perfectly compatible with some downstream applications, such as ControlNet, which typically rely on 1D conditions. Recently, Make-An-Audio-2 [10] introduces a 1D Variational Autoencoder (VAE) for mel-spectrograms, utilizing a transformer-based architecture that has demonstrated superior generation quality compared to 2D representations. However, as noted in [11], [12], the reconstruction of mel spectrograms might still lead to degraded audio quality, particularly for sound effects and music. To tackle these challenges and integrate DiT into T2A, we utilize a waveform VAE [13], [14], which reduces computational costs, preserves high temporal resolution, and eliminates the need for an additional neural vocoder while delivering strong reconstructions.

*This work was done while J. Hai was an intern at Tencent AI lab, USA.

Combining the characteristics of waveform latents, the prediction target in diffusion modeling, and the conditioning method in T2A, we carefully redesign the DiT by introducing a novel adaptive layer norm (AdaLN) method, incorporating long-skip connections, and leveraging techniques like RoPE [15] and QK-Norm [16]. The proposed EzAudio-DiT achieves fast convergence and stable training while using fewer parameters and reducing memory consumption.

Another major challenge in T2A generation is the lack of large-scale annotated datasets. AudioLDM [4] relies on CLAP [17] embeddings for unlabeled audio but struggles with generation performance due to mismatches between text and audio embeddings. Make-an-Audio-1&2 [6], [10] propose using synthetic audio data, but the dataset is not fully open-sourced, and synthesizing new data could be time-consuming. Also, synthetic data introduces noticeable artifacts due to inconsistent recording environments of sound samples, limiting the quality of audio generation. Tango [2] introduces TangoPrompt-Bank, a collection of synthetic caption datasets, but the inconsistent quality and time-consuming organization of multiple datasets remain barriers. To address these challenges, we propose a three-stage training pipeline using open-sourced datasets: Audioset [18], VGGSound [19], and AudioCaps [20]. First, we integrate masked modeling to let the model learn acoustic dependency. Next, we use audio caption data automatically generated and refined by audio-language models and language models for text-audio alignment training. Finally, we fine-tune the model on human-labeled data for precise audio generation. Our strategy enhances both generation quality and prompt alignment. The generated captions are released, and with the audio data being open-sourced, they are easily accessible for future research.

In addition, classifier-free guidance (CFG) [21] is widely used in diffusion model sampling, where increasing the CFG score can enhance prompt alignment but may degrade generation quality due to over-exposure. This problem is particularly pronounced with latent waveform representations, as waveform peak distributions affect not only loudness and dynamic range but also frequency characteristics. To address this issue, we incorporate a CFG rescaling technique [22]. This approach ensures strong prompt alignment with a negligible compromise to audio fidelity when using higher CFG scores. As a result, it eliminates the need to carefully balance CFG scores, simplifying the use of the model.

In summary, we introduce EzAudio, an innovative and *easy-to-follow* T2A framework that (1) operates in waveform latent space, (2) features a newly designed efficient T2A DiT architecture, (3) incorporates a novel training strategy, and (4) enhances CFG during diffusion sampling. EzAudio produces highly realistic audio samples, outperforming existing open-source models in both objective and subjective evaluations. The code, dataset, and model checkpoints are released to help researchers and startups build T2A models *easily* and *efficiently*. We also hope EzAudio can inspire advancements in other audio-generation tasks such as video-to-audio synthesis and beyond.

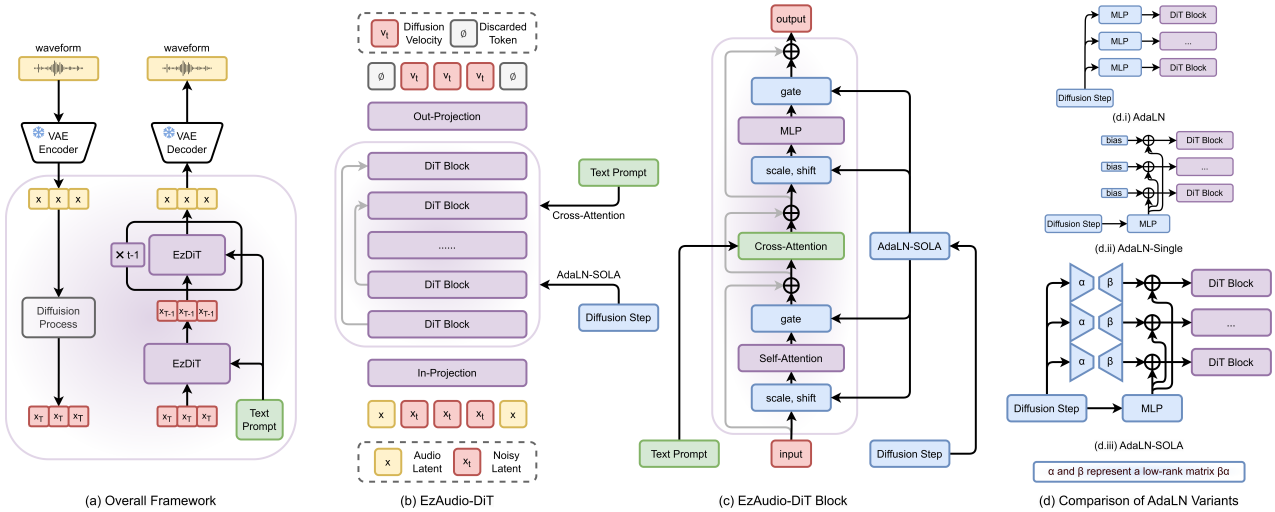

Fig. 1. The framework of the proposed EzAudio and the architectural details of EzAudio-DiT.

II. Method

In this section, we introduce the key components and techniques that underpin EzAudio, including the architecture design, classifier-free guidance rescaling, and the multi-stage training strategy.

A. Overview of EzAudio

EzAudio is composed of three key components: (1) a text encoder, (2) a latent diffusion model (LDM), and (3) a waveform VAE. The text encoder processes the input audio description, which is then used by the latent diffusion model to generate a latent representation of the audio waveform, starting from standard Gaussian noise through reverse diffusion. Finally, the waveform VAE decoder reconstructs this latent representation into an audio waveform.

EzAudio uses FLAN-T5 [23], which has demonstrated great performance in T2A tasks [2], [3], as the text encoder. The LDM employs velocity (v) prediction [24] and Zero-SNR diffusion schedulers [22], both of which have been successful in diffusion-based image and audio generation [25], [26]. The neural network in LDM is based on EzAudio-DiT, a transformer model specifically designed for T2A.

The waveform VAE is based on Stable Audio [27] and DAC [28], utilizing a fully convolutional autoencoder with snake activation functions [29], but with a VAE bottleneck instead of residual vector quantization (RVQ). The VAE is trained with a combination of KL divergence, reconstruction, and GAN losses to ensure a Gaussian-distributed latent space and high-quality audio reconstructions. We train[1] the waveform VAE on AudioSet [18], enabling it to handle a wide variety types of audio.

B. Proposed Efficient EzAudio-DiT

EzAudio proposes several innovative designs to make DiT better suited for T2A by improving parameter and memory efficiency, convergence speed, and training stability. These innovations include:

- **AdaLN-SOLA:** The AdaLN layers in DiT account for a significant portion of the parameters, as they manage both image class conditions and diffusion steps. However, with cross-attention handling the text inputs, the task managed by AdaLN becomes simpler, making it straightforward to simplify AdaLN. AdaLN-single, introduced by [9], aims to reduce the model size and memory usage in DiT with cross-attention by using a single shared AdaLN across all DiT blocks. However, we find that AdaLN-Single leads to performance degradation and makes DiT training unstable. To address these issues, we propose AdaLN-SOLA (**AdaLN-S**ingle **O**rchestrated by **L**ow-rank **A**djustment). As shown in Fig. 2.d.iii, AdaLN-SOLA uses one shared AdaLN module, but each block uses a low-rank matrix that takes the diffusion step as input to adaptively adjust the shared AdaLN. Thus, while reducing model parameters and memory usage, it can still maintain model performance and numerical stability.
- **Long-skip Connection:** In diffusion models, the input low-level features contain essential information for accurate noise or velocity estimation. When working with waveform latent embeddings, which have 128 channels—far more than typical image representations—the transformer struggles to retain detailed input information. To alleviate this burden, we apply long-skip connections that create shortcuts for low-level features to reach the later blocks in the transformer, as shown in Fig. 1.b.
- **Other Techniques:** To stabilize DiT during training, we apply QK-Norm [16] in the attention layers and introduce LayerNorm [30] after the fusion of long-skip connections. Additionally, we incorporate RoPE [15], which has been shown to accelerate transformer convergence and improve model performance, to handle the position encoding of audio latents.

C. Pre-training via Masked Modeling and Synthetic Captions

Compared to T2I, T2A faces the issue of insufficient data. To unlock the potential of the diffusion transformer and improve the model's performance, we adopted a multi-stage training approach like [31] and [9], which consists of the following stages:

1) Masked Diffusion Modeling: Masked modeling has been successfully applied in transformers [32], [33] and diffusion transformers [34] for efficient self-supervised pre-training. In this stage, we utilize AudioSet [18], a large-scale dataset with diverse audio classes but noisy annotations. During training, a random portion of tokens, ranging from 25% to 100% with a minimum span of 0.2s, is masked by adding diffusion noise. The model is then trained to reconstruct the masked tokens using the unmasked ones, without text conditioning. When fully masked, the model operates as an unconditional model.

2) Synthetic Caption Data Generation: Next, we utilize synthetic caption data for text-audio alignment training. To increase audio and language diversity, we incorporate multiple sources of synthetic data:

[1]VAE Training details are illustrated on EzAudio's website.

- **Auto-ACD [35]**: An open-source dataset with 1.5 million captions for AudioSet and VggSound. During caption generation, audio and video caption models produce initial captions, which are then refined by a language model into natural audio captions.
- **AS-Qwen-Caps**: Audio captions generated for AudioSet using Qwen-Audio[2] [36], one of the leading audio-language models.
- **AS-SL-GPT4-Caps**: Audio captions created using OpenAI's GPT-4o-mini API[3], based on temporal annotations from the strongly labeled subset of AudioSet.

To ensure high-quality captions, we use a filtering method similar to CapFilt [38]. Using a pre-trained CLAP model [17], we filter out audio-caption pairs with similarity scores below a set threshold.

Building on the model from the first stage, we incorporate a cross-attention module into each DiT block to process text conditions. To ensure stable training resumption, we initialize the output projection layer of the cross-attention module to zero. Additionally, we raise the likelihood of applying a full mask, to encourage the model to rely more on the text input. Furthermore, 10% of the text is replaced with empty input to enable unconditional modeling for CFG.

3) Fine-tuning: Finally, following the approach used in Tango, we fine-tune the model on AudioCaps [20], a manually labeled audio caption dataset, ensuring accurate and high-quality audio generation.

D. Improving Sampling with Classifier-Free Guidance Rescaling

The CFG [21] is utilized 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 may result in over-exposure, impairing generation quality. To address this, a CFG rescaling technique [22] 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.

III. Experiments

A. Experimental Setups

We conducted experiments using a 24kHz audio sample rate for both the waveform VAE and the T2A model. The waveform latent representation operates at 50Hz and consists of 128 channels. For DiT variants, DiT-L consists of 24 DiT blocks, each with 1024 channels, while DiT-XL has 28 DiT blocks, each with 1152 channels. All the models are trained with the AdamW optimizer. During diffusion sampling, we use 50 steps and a CFG score of 3 by default.

Following previous T2A studies [2], [4], [5], [10], we evaluate our model using Frechet Distance (FD)[4], Kullback–Leibler (KL) divergence, and Inception Score (IS), with pre-trained PANNs [39] as the feature extractor. Additionally, we employ CLAP[5] [17] to assess the alignment between the generated audio and the text prompt. All audio samples are resampled to 16kHz before evaluation.

The AudioCaps test set, comprising 900 audio clips with 882 currently available, is used for evaluation. Each clip has five human-written captions, and we randomly select one caption[6] per clip, following the approach used in [2], [4].

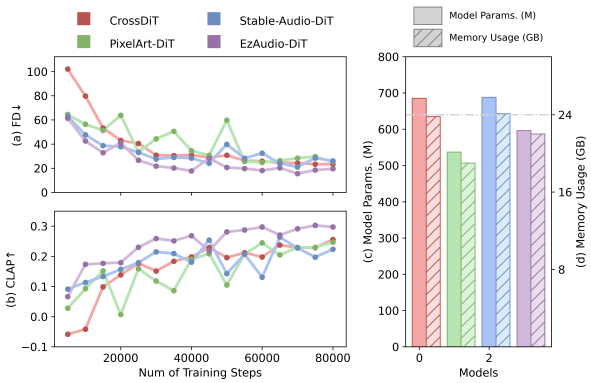

Fig. 2. Comparison of diffusion transformer architectures.

B. Comparison of DiT Architecture in T2A

We perform an ablation study on different DiT variants using the AudioCaps dataset, training for 80k steps with a batch size of 128 and a learning rate of 1e-4. The configuration for the number of blocks and transformer channels follows the DiT-L setup outlined in Section III-A. The variants include CrossDiT, which adds cross-attention layers to the vanilla DiT [7], Pixel-Art-DiT [9], which replaces AdaLN with AdaLN-Single, Stable-Audio-DiT [27], designed for text-to-music generation and incorporating RoPE and QK-Norm, and the proposed EzAudio-DiT, detailed in Section II-B.

As shown in Fig. 2.a, PixelArt-DiT converges quickly during the early stages but becomes unstable with more training steps. With the help of RoPE, Stable-Audio-DiT converges faster than CrossDiT initially, but the two models' performance becomes comparable later, with Stable-Audio-DiT being less stable. The proposed EzAudio-DiT shows fast and consistent convergence among the DiT variants, eventually achieving better performance. We attribute[7] the faster convergence to the long-skip connections and RoPE, while the stable training is a result of AdaLN-SOLA, QK-Norm, and Skip-Norm.

In Fig. 2, we compare the model parameter and memory usage with a training batch size of 16. PixelArt-DiT, using AdaLN-Single, has the smallest parameter count and lowest memory usage. EzAudio-DiT, featuring AdaLN-SOLA and skip connections, shows slightly higher parameters and memory usage than PixelArt-DiT but remains significantly lower than Cross-DiT and Stable-Audio-DiT. Notably, with a batch size of 16, EzAudio-DiT's memory usage stays under 24GB, enabling efficient fine-tuning on GPUs like the Nvidia 4090.

C. Comparison of Training Methods

In this section, we compare different pre-training strategies that utilize larger datasets, employing the DiT-XL configuration for EzAudio-DiT, as described in Section III-A. The training process is divided into three stages, all with a batch size of 128. Stage 1 uses 1.80M samples over 100K training steps with a learning rate of 1e-4.

[2]We compare Qwen-Audio [36] and GAMA [37], selecting Qwen-Audio for its higher accuracy and fewer hallucinations on AudioCaps evaluation.

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

[4]We exclude FAD due to reliability concerns raised in prior works [3], [4].

[5]We use the latest version: https://huggingface.co/laion/larger_clap_general

[6]Some studies [10] use all captions for evaluation, leading to a lower FD.

[7]Additional ablation studies can be found on EzAudio's website.

TABLE I
EVALUATION RESULTS WITH DIFFERENT PRE-TRAINING METHODS.

Strategy	Threshold	FD↓	KL↓	IS↑	CLAP↑
Tango's [2]	/	17.79	1.66	9.60	0.273
EzAudio's	0.35	16.17	1.48	9.85	0.290
EzAudio's	0.40	**15.46**	1.44	10.11	0.294
EzAudio's	0.45	16.27	**1.40**	**10.31**	**0.303**

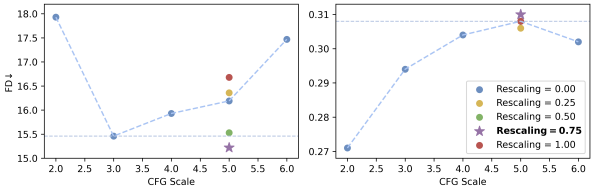

Fig. 3. FD and CLAP scores across CFG scales and rescaling factors.

In Stage 2, 50K steps are performed with a learning rate of 5e-5, using 0.58M, 0.27M, or 0.11M samples depending on thresholds of 0.35, 0.40, and 0.45, respectively. Stage 3 completes the training with 30K steps, using 48K samples and a learning rate of 1e-5. The entire training process takes 5 days using 8 A100-40G GPUs.

Table I compares our proposed training strategy with Tango [2], which uses the TangoPromptBank, a collection of audio caption datasets, for pre-training, and AudioCaps for fine-tuning. For Tango-PromptBank, we use a batch size of 128, with 150K steps at a learning rate of 1e-4, followed by 30K fine-tuning steps at a learning rate of 1e-5. Our method achieves superior generation quality and stronger text-audio alignment compared to Tango's. Additionally, we evaluate different thresholds for filtering synthetic captions: a lower threshold allows for more diverse but noisier data, negatively impacting all metrics, while a higher threshold improves most metrics but reduces FD and data diversity. We select a threshold of 0.40, as it provides the best balance between data diversity and model performance.

D. Impact of CFG and CFG Rescaling

As shown in Fig. 3, higher CFG values enhance text-audio alignment but increase FD, indicating reduced audio quality. Since a CFG of 5 yields the highest CLAP score and minor degradation in FD, we apply rescaling at this level. Using a rescaling factor of around 0.50–0.75 allows the model to maintain strong text alignment while mitigating the negative effects on audio quality.

E. Comparison with State-of-the-art

We compare EzAudio-L and EzAudio-XL, both trained using the proposed strategy but with different DiT configurations, as described in Section III-A, to recent open-source T2A models. Tango[8]-1&AF[9] [2], [3], AudioLDM-1&2[10] [4], [5], and Make-An-Audio [6], all employ a 2D U-Net-based diffusion approach with mel-spectrogram audio representations. AudioLDM and Make-An-Audio use CLAP as their text encoder, while Tango-1 and Tango-AF use FLAN-T5. AudioLDM-2 introduces a GPT-2-based encoder that works with both CLAP and FLAN-T5 inputs. Make-An-Audio-2 [10], using a 1D-VAE for mel-spectrogram representation, adopts a transformer architecture, utilizing CLAP for original text prompts and a fine-tuned FLAN-T5 for GPT-3.5-processed prompts. For a fair comparison, we used official checkpoints from each model's public repository and evaluated them on the AudioCaps testset as outlined in Section III-A. We follow the recommended sampling steps and CFG score settings from each method's paper or repository. Specifically, we use 200 steps for Tango-1&AF and AudioLDM-1&2, and 100 steps for Make-An-Audio-1&2. For EzAudio, we apply 100 steps with a CFG score of 5 with a rescaling factor of 0.75.

[8]Since our model does not use Direct Preference Optimization (DPO) [40], we leave a comparison with Tango-2 for future work.

[9]We use *tango-full-ft-audiocaps* and *tango-af-ac-ft-ac* from Tango's repo.

[10]We use *audioldm-l-full* and *audioldm2-large* from cvssp's repo.

TABLE II
THE COMPARISON BETWEEN EZAUDIO AND T2A MODELS ON THE AUDIOCAPS DATASET. †INDICATES TRAINABLE PARAMETERS.

Model	# Params.†	FD↓	KL↓	IS↑	CLAP↑
Ground Truth	–	–	–	–	0.302
Tango [2]	866M	19.07	1.33	7.70	0.293
Tango-AF [3]	866M	21.84	1.32	9.20	0.269
AudioLDM-Large [4]	739M	30.96	2.36	7.38	0.197
AudioLDM-2-Large [5]	712M	25.03	1.75	8.13	0.236
Make-An-Audio [6]	453M	18.77	1.71	8.80	0.244
Make-An-Audio-2 [10]	937M	16.16	1.42	9.93	0.284
EzAudio-L	596M	15.59	1.38	11.35	**0.319**
EzAudio-XL	874M	**14.98**	**1.29**	**11.38**	0.314

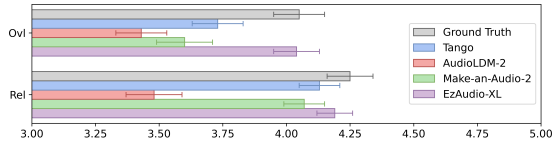

Fig. 4. Mean subjective scores with 95% confidence intervals.

As shown in Table II, AudioLDM-1&2 exhibit weaker overall performance compared to the other baselines. While Make-An-Audio 2 achieves higher FD and IS scores, it falls slightly behind in text-audio alignment, likely due to its reliance on a language model to categorize sound events into broad segments. Tango performs well in alignment but produces less realistic audio, and although Tango-AF improves IS, it underperforms in FD and CLAP scores. In contrast, both EzAudio-L and EzAudio-XL outperform the baseline methods in terms of quality and text-audio alignment, with EzAudio-XL showing a slight edge over EzAudio-L across most metrics.

We conducted a subjective experiment to evaluate overall audio quality (OVL) and text prompt relevance (REL) using a 5-point Mean Opinion Score (MOS) on 30 randomly selected text prompts. Twelve participants with backgrounds in music production or audio engineering took part in the experiment. We compare EzAudio-XL with Tango, AudioLDM-2, Make-an-Audio-2, and audio samples from AudioCaps. As shown in Fig. 4, the results are consistent with objective findings: EzAudio-XL outperforms the baselines in both text alignment and audio quality. While Make-an-Audio 2 shows higher FD and IS scores than Tango, it occasionally produces artifacts, likely due to the use of synthetic data. Notably, EzAudio-XL's OVL score approaches real recordings, highlighting its ability to generate realistic audio.

IV. CONCLUSION

In this paper, we introduce EzAudio, a novel *easy-to-deploy* and *easy-to-use* T2A framework. EzAudio leverages an efficient DiT architecture, a streamlined training pipeline with synthetic caption data, and a CFG rescaling technique to achieve both precise and high-quality audio generation. The proposed framework generates highly realistic audio and achieves state-of-the-art performance. In the future, we plan to integrate ControlNet and DreamBooth and further explore the applications of EzAudio in voice and music generation.

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] P. P. Li, B. Chen, Y. Yao, Y. Wang, A. Wang, and A. Wang, "Jen-1: Text-guided universal music generation with omnidirectional diffusion models," in *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024, pp. 762–769.

[13] G. L. Lan, B. Shi, Z. Ni, S. Srinivasan, A. Kumar, B. Ellis, D. Kant, V. Nagaraja, E. Chang, W.-N. Hsu *et al.*, "High fidelity text-guided music generation and editing via single-stage flow matching," *arXiv preprint arXiv:2407.03648*, 2024.

[14] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," *arXiv preprint arXiv:2402.04825*, 2024.

[15] 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.

[16] A. Henry, P. R. Dachapally, S. Pawar, and Y. Chen, "Query-key normalization for transformers," *arXiv preprint arXiv:2010.04245*, 2020.

[17] 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.

[18] 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.

[19] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[20] 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.

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

[22] 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.

[23] 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.

[24] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[25] 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.

[26] 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.

[27] 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.

[28] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[29] L. Ziyin, T. Hartwig, and M. Ueda, "Neural networks fail to learn periodic functions and how to fix it," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1583–1594, 2020.

[30] J. Ba, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[31] 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.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[33] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.