MeanAudio: Fast and Faithful Text-to-Audio Generation with Mean Flows

Xiquan Li[1,2†], **Junxi Liu**[1†], **Yuzhe Liang**[1], **Zhikang Niu**[1], **Wenxi Chen**[1], **Xie Chen**[1‡]

[1]MoE Key Lab of Artificial Intelligence, X-LANCE Lab
Department of Computer Science, Shanghai Jiao Tong University, China
[2]SJTU Paris Elite Institute of Technology, Shanghai Jiao Tong University, China
{mtxiaoxi55, chenxie95}@sjtu.edu.cn

Abstract

Recent years have witnessed remarkable progress in Text-to-Audio Generation (TTA), providing sound creators with powerful tools to transform inspirations into vivid audio. Yet despite these advances, current TTA systems often suffer from slow inference speed, which greatly hinders the efficiency and smoothness of audio creation. In this paper, we present MeanAudio, a fast and faithful text-to-audio generator capable of rendering realistic sound with only one function evaluation (1-NFE). MeanAudio leverages: (i) the MeanFlow objective with guided velocity target that significantly accelerates inference speed, (ii) an enhanced Flux-style transformer with dual text encoders for better semantic alignment and synthesis quality, and (iii) an efficient instantaneous-to-mean curriculum that speeds up convergence and enables training on consumer-grade GPUs. Through a comprehensive evaluation study, we demonstrate that MeanAudio achieves state-of-the-art performance in single-step audio generation. Specifically, it achieves a real-time factor (RTF) of 0.013 on a single NVIDIA RTX 3090, yielding a 100x speedup over SOTA diffusion-based TTA systems. Moreover, MeanAudio also shows strong performance in multi-step generation, enabling smooth transitions across successive synthesis steps.

1 Introduction

Text-to-Audio Generation (TTA) (Liu et al., 2023a; Ghosal et al., 2023; Huang et al., 2023b) aims to synthesize diverse auditory content from textual prompts. By translating language into sound, TTA models unlock a broad spectrum of real-world applications, including virtual reality, gaming, film post-production, and human-computer interaction.

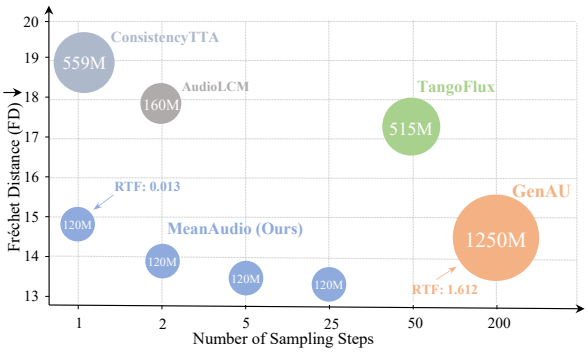

Figure 1: MeanAudio achieves state-of-the-art single-step generation performance with a real-time factor (RTF) of 0.013, offering a 100x speedup over existing diffusion-based TTA systems. It also demonstrates strong performance in multi-step generation, despite using only 120M parameters.

In pursuit of stronger audio generators, recent studies have advanced along three complementary directions: architectural innovations (Huang et al., 2023a; Hai et al., 2025; Hung et al., 2024; Haji-Ali et al., 2024; Evans et al., 2025; Lee et al., 2024; Valle et al., 2025), scaling data and model parameters (Kong et al., 2024; Yuan et al., 2025; Liu et al., 2024a; Haji-Ali et al., 2024), and employing novel training objectives (Majumder et al., 2024; Liao et al., 2024). These approaches have substantially improved generation quality and controllability, as evidenced by the Inception Score (IS) on AudioCaps (Kim et al., 2019) rising from 8.17 (Liu et al., 2023a) to 12.81 (Hung et al., 2024). However, despite these improvements, current TTA models often suffer from slow inference speed, where they typically require seconds to minutes to synthesize a single audio sample. Such latency, stemming from the iterative sampling process in flow and diffusion-based models, not only impedes deployment in time-sensitive scenarios such as virtual assistants and interactive gaming, but also hinders the creative workflow of sound creators.

To accelerate the inference speed of TTA models, recent studies (Liu et al., 2024b, 2025; Saito et al.,

[†]Equal contributions, [‡]Corresponding author.
[‡]Code: https://github.com/xiquan-li/MeanAudio
Demo: https://MeanAudio.github.io/

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14378–14393

2025; Bai et al., 2023) have primarily focused on diffusion distillation (Song et al., 2023). In this paradigm, the number of diffusion sampling steps is reduced by distilling a pre-trained teacher model into a few-step student generator. As such, the student generator learns to synthesize audio in a few steps by modeling the flow trajectory of their multi-step teachers. While these models have achieved promising performance in single-step and few-step audio generation, they are inherently limited by the rigid consistency constraints and their reliance on teacher models. Moreover, the distillation-based approaches are often computationally expensive, as online methods require holding 2-3 full models in memory simultaneously, and offline methods rely on large-scale generation and storage of teacher trajectories before training.

In this paper, we present MeanAudio, a MeanFlow-based (Geng et al., 2025) fast and faithful text-to-audio generator capable of rendering high-quality audio with only 1 iteration. To improve synthesis quality, MeanAudio leverages an enhanced Flux-Style (BlackForestLabs, 2024) flow transformer with dual text encoders, facilitating realistic and instruction-adherent audio generation. To accelerate inference, MeanAudio regresses the average velocity field during training, enabling direct mapping from the start to the endpoint of the flow trajectory. By further integrating classifier-free guidance (CFG) (Ho and Salimans, 2022) into the training objective, it achieves guided sampling without additional computational cost. Moreover, we introduce an instantaneous-to-mean learning curriculum with flow-field mix-up, which anchors the model in the instantaneous velocity field before progressively adapting to the average velocity field. This strategy proves beneficial for improving both training efficiency and generation performance across single- and multi-step inference.

Through extensive experiments, we show that MeanAudio achieves state-of-the-art (SOTA) performance in single-step TTA generation. Notably, it achieves a real-time factor (RTF) of 0.013, corresponding to 100x speedup over the best open-sourced diffusion-based TTA system, GenAU (Haji-Ali et al., 2024), which requires 200 sampling steps. Beyond single-step generation, MeanAudio also demonstrates competitive performance in multi-step synthesis, all within a compact 120M-parameter model that can be efficiently trained in three days on four NVIDIA RTX 3090 GPUs. In addition, to uncover best practices for building MeanFlow-accelerated audio generators, we conduct a comprehensive ablation study that highlights the importance of architectural choices, training strategies, and flow configurations. To summarize, our main contributions are as follows:

- We present **MeanAudio**, the first text-to-audio generator that learns MeanFlows for fast and faithful sound synthesis.
- We design an improved Flux-Style flow transformer with dual text encoders, enabling high-quality and prompt-adherent audio generation.
- We introduce an instantaneous-to-mean curriculum with flow mix-up that facilitates stable training and rapid convergence.
- Extensive experiments show that MeanAudio achieves SOTA results in single-step generation and competitive performance in multi-step generation.
- We provide comprehensive experimental analysis that highlights best practices for building MeanFlow-based audio generators, paving the way toward faster and stronger audio models.

We will fully release the MeanAudio codebase and model weights to facilitate future research on efficient and high-quality text-to-audio generation.

2 Preliminaries

2.1 Conditional Flow Matching

Flow Matching (Liu et al., 2023b; Lipman et al., 2023; Albergo and Vanden-Eijnden, 2023) is a powerful generative model that learns to match the flows between two probabilistic distributions. Given data $x \sim p_{\text{data}}(x)$, prior $\epsilon \sim p_{\text{prior}}(\epsilon)$, the optimal transport flow path can be constructed as: $x_t = (1-t)x + t\epsilon$, and the conditional velocity is thus given by $v_t = \frac{dx_t}{dt} = \epsilon - x$. At training time, the objective is to find a neural network f_θ which minimizes the conditional flow matching loss (Lipman et al., 2023):

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,x,\epsilon} \| f_\theta(t, x_t) - v_t \|^2 \tag{1}$$

During sampling, we randomly draw noise ϵ from p_{prior} and solve the ordinary differential equation (ODE) defined below:

$$dx_t = -f_\theta(t, x_t)dt$$

The solution is thus given by $x_r = x_t - \int_r^t f_\theta(\tau, x_\tau)d\tau$, where r denotes another time step. During implementation, we can use numerical methods (e.g. Euler Method) to approximate this integration.

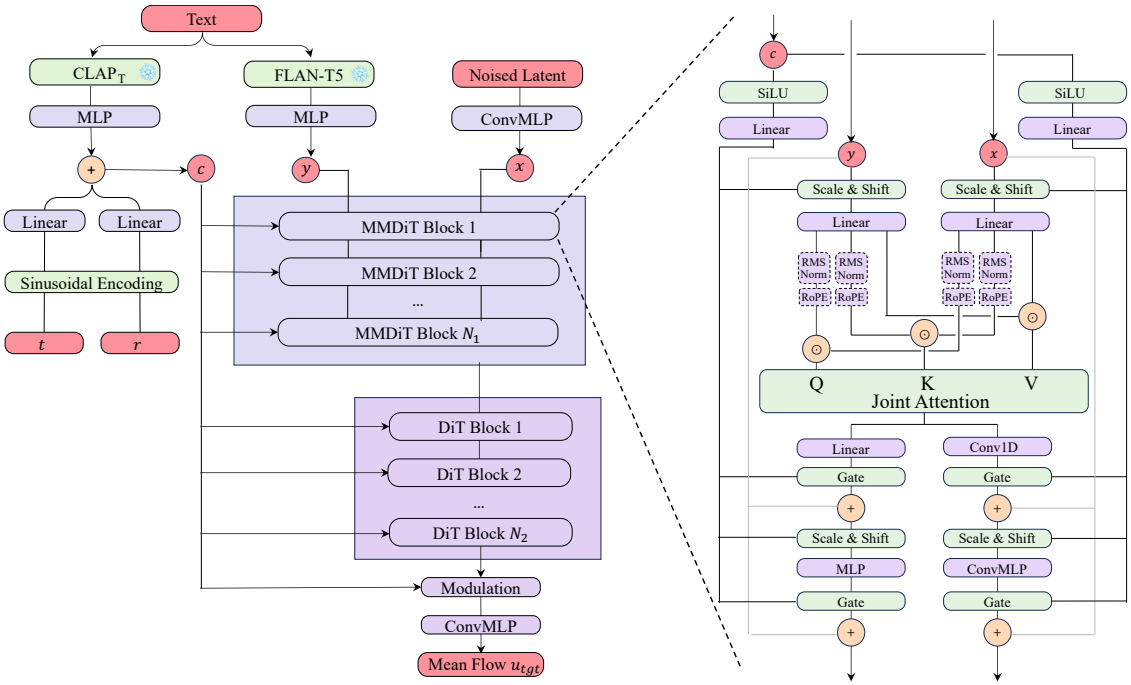

Figure 2: Model architecture overview: MeanAudio combines N_1 multi-modal (MMDiT) blocks and N_2 single-modal (DiT) blocks to construct the flow transformer. It leverages the joint attention in multi-modal blocks to integrate FLAN-T5's fine-grained text embeddings, and employs AdaLN to inject CLAP's global conditioning.

2.2 Mean Flows for Generative Modeling

To accelerate the inference speed of Flow Matching, Mean Flows (Geng et al., 2025) proposed to regress the average velocity field during training, allowing high-quality single-step generation. Specifically, given a time interval $[r, t]$, the average velocity within it is defined as: $u(x_t, r, t) \triangleq \frac{1}{t-r}\int_r^t v(x_\tau, \tau)d\tau$. By differentiating both sides with respect to t and re-arranging terms, we obtain the *Mean Flow Identity*, which describes the relation between v and u:

$$u(x_t, r, t) = v_t - (t-r)\frac{d}{dt}u(x_t, r, t)$$

We then encourage f_θ to satisfy this identity by minimizing the mean flow objective:

$$\mathcal{L}_{\text{MF}} = \mathbb{E}_{t,r,x,\epsilon}\|f_\theta(x_t, r, t) - \text{sg}(u_{\text{tgt}})\|^2 \quad (2)$$

Where $u_{\text{tgt}} = v_t - (t-r)\frac{d}{dt}f_\theta(x_t, r, t)$, and $\text{sg}(\cdot)$ denotes the stop-gradient operation. Note that this total derivative can be expanded by its partial components, corresponding to a Jacobian-Vector Product (JVP): $\frac{d}{dt}f_\theta(x_t, r, t) = v_t\partial_x f_\theta + \partial_t f_\theta$. When $r = t$, the mean flow objective becomes the vanilla flow matching objective.

During sampling, the time integral in CFM can be replaced by the average velocity, leading to:

$$x_r = x_t - (t-r)f_\theta(x_t, r, t)$$

In particular, in single-step generation, we have: $x_0 = x_1 - f_\theta(x_1, 0, 1)$ where $x_1 = \epsilon \sim p_{\text{prior}}(\epsilon)$.

3 MeanAudio

As illustrated in Figure 2, MeanAudio employs an enhanced Flux-Style flow transformer to learn average velocity in the latent space conditioned on the textual prompt and timestep embeddings. In this section, we first describe the architectural design of MeanAudio, and then introduce our strategies for accelerating training and inference.

3.1 Audio Encoding

Following prior works (Liu et al., 2023a; Ghosal et al., 2023), we model the generative process in the latent space to improve computational efficiency. Specifically, we apply the short-time Fourier transform (STFT) to audio waveforms and extract the magnitude component as mel spectrograms (Stevens et al., 1937). These spectrograms are then encoded into latent representations x using a pre-trained variational autoencoder (VAE) (Kingma and Welling, 2013). During inference, the generated latents are decoded back into spectrograms via the VAE and subsequently converted to audio waveforms using a pre-trained vocoder (Lee et al., 2023). We employ a 1D convolution-based VAE for its superior capacity to model frequency- and length-variable representations. For a 10-second audio input, the autoencoder produces a latent sequence of 312 tokens, each with a hidden dimension of 20.

3.2 Enhanced Flow Transformer

We design an enhanced Flux-Style flow transformer to improve MeanAudio's synthesis quality and prompt adherence. Specifically, we combine N_1 multi-modal transformer blocks (MMDiT) (Esser et al., 2024) with audio/text branches and N_2 audio-only DiT (Peebles and Xie, 2023) blocks to construct our flow transformer. To further boost generation quality, we employ several refinements: Firstly, we use ConvMLP rather than vanilla MLPs in the audio stream of MeanAudio. ConvMLP uses 1D convolutions (kernel size = 3 and padding = 1) rather than linear layers, demonstrating stronger performance in capturing local temporal structure (Cheng et al., 2024). Secondly, we apply rotary positional embedding (RoPE) (Su et al., 2024) on the queries and keys in both the audio and text branches. Unlike absolute position embeddings, RoPE models the relative distances and is beneficial for variable-length audio generation. Thirdly, we use RMSNorm (Zhang and Sennrich, 2019) with learnable scales in attention calculation to enable stable and efficient training. Lastly, we adopt the SwiGLU (Shazeer, 2020) activations instead of ReLU (Agarap, 2018) in the MLP layers.

3.3 Model Conditioning

MeanAudio is conditioned on a textual prompt and time steps to render faithful audio signals. For textual conditioning, we leverage FLAN-T5 (Chung et al., 2024) and CLAP (Wu et al., 2023) to extract caption embeddings. FLAN-T5 is an instruction-tuned large-language model (LLM) capable of producing fine-grained token embeddings. Meanwhile, CLAP is pre-trained on large-scale audio-text dataset and can offer global acoustic-aligned text embeddings. Denote $y_{\text{T5}} \in \mathbb{R}^{N \times d_{\text{T5}}}$ as the embedding extracted by FLAN-T5, where N and d_{T5} represent the number of tokens and model's output dimension. We feed y_{T5} into the text branch of MMDiT, where multi-modal joint attention learns cross-modal alignment between text and audio. Furthermore, let $y_{\text{CLAP}} \in \mathbb{R}^{1 \times d_{\text{CLAP}}}$ represent the global text embedding obtained from the CLAP text encoder (CLAP_{T}). We project this embedding through an MLP and combine it with the extracted timestep features to form the global condition $c = t_{\text{emb}} + r_{\text{emb}} + y'_{\text{CLAP}}$. This global information is then injected into the model via the scales and biases of adaptive layer normalization (AdaLN) layers. While FLAN-T5 can capture fine-grained textual details at the token level, CLAP contributes holistic, audio-grounded semantic information. Together, they provide rich and balanced conditions that improve both the fidelity and semantic alignment of the generated audio.

3.4 Integrated Classifier-Free Guidance

Classifier-free guidance (CFG) (Ho and Salimans, 2022) is a widely adopted technique for achieving controllable generation. However, using CFG during sampling doubles the number of function evaluations (NFE), as both class-conditional and unconditional model outputs should be computed. To eliminate the additional cost associated with guided sampling, MeanAudio integrates CFG into the training target. Specifically, define v_t^{cfg} as the estimated instantaneous velocity field with guidance, which can be expressed as follows:

$$v_t^{\text{cfg}} = \omega v_t + \underbrace{\kappa f_\theta(x_t, t, t|\mathbf{C})}_{\text{cls. conditional}} + \underbrace{(1-\omega-\kappa) f_\theta(x_t, t, t|\varnothing)}_{\text{cls. unconditional}} \quad (3)$$

Here, κ is a mixing factor which combines both class-conditional and unconditional predictions into the guided field, resulting in an effective guidance scale of $\omega' = \frac{\omega}{1-\kappa}$. Similarly, we also expose the trainable network f_θ with class-unconditional inputs, where we randomly drop $\mathbf{C}$ with 10% probability, following (Ho and Salimans, 2022). By replacing v_t with v_t^{cfg} in u_{tgt}, we obtain the average velocity target with guidance, which can be formulated as:

$$u_{\text{tgt}}^{\text{cfg}} = v_t^{\text{cfg}} - (t-r)\frac{d}{dt} f_\theta(x_t, r, t)$$

By regressing $u_{\text{tgt}}^{\text{cfg}}$, MeanAudio directly learns the guidance during training, thus avoiding the need for an additional forward pass during generation. The training and inference procedure of MeanAudio are illustrated in Algorithm 1 and 2.

3.5 Stabilizing Flow Fields

Although the *Mean Flow Identity* provides an effective training target for learning fast single-step generation, we found that directly modeling audio latent with Eq. 2 results in unstable training, slow convergence, and poor multi-step generation. This may be due to several factors: Firstly, the MeanFlow objective defined in Eq. 2 focuses solely on learning the average velocity, which may cause the

Algorithm 1 MeanAudio Training

Input: Encoded audio latent: x, textual conditions $\mathbf{C}$, flow transformer f_θ.
Sample $t, r \sim \text{lognorm}(\mu, \sigma)$, $\epsilon \sim \mathcal{N}(0, I)$
$x_t \leftarrow (1-t) \cdot x + t \cdot \epsilon$
$v_t \leftarrow \epsilon - x$
Compute model output $f_\theta(x_t, r, t)$ and its derivative $\frac{df_\theta}{dt}$ via JVP:
$f_\theta(x_t, r, t), \frac{df_\theta}{dt} \leftarrow \text{JVP}(f_\theta, (x_t, r, t), (v, 0, 1))$
Compute the guided instantaneous velocity estimate:
$v_t^{\text{cfg}} \leftarrow \omega \cdot v_t + \kappa \cdot f_\theta(x_t, t, t, \varnothing) + (1 - w - \kappa) \cdot f_\theta(x_t, t, t, \mathbf{C})$
Compute the guided average velocity estimate:
$u_{\text{tgt}}^{\text{cfg}} \leftarrow v_t^{\text{cfg}} - (t - r) \cdot \frac{d}{dt} f_\theta$
loss $\leftarrow \|f_\theta(x_t, r, t) - \text{sg}(u_{\text{tgt}})\|^2$
Output: loss

Algorithm 2 MeanAudio Inference

Input: Trained network f_θ, textual conditions $\mathbf{C}$
Sample noise: $\epsilon \sim \mathcal{N}(0, I)$
$x \leftarrow \epsilon$
for $i = 0$ to $N - 1$ **do**
    $x = x - (t_{i+1} - t_i) \cdot f_\theta(x, t_i, t_{i+1}, \mathbf{C})$
end for
Output: x

model to neglect the underlying instantaneous field that serves as the foundation for Mean Flow (Peng et al., 2025). Secondly, the training target $u_{\text{tgt}}^{\text{cfg}}$ is defined by the model's own derivative. However, a randomly initialized model may fail to provide effective guidance at the beginning, resulting in slow convergence.

To address this issue, we propose an instantaneous-to-mean curriculum with flow field mixup to improve training stability and efficiency. As illustrated in Figure 3, our curriculum comprises two stages: In the first stage, the model is trained on large-scale, weakly-labeled audio-text datasets to learn the instantaneous velocity field, whose loss is defined in Equation 1. This stage establishes a strong initialization, allowing the model to first capture the underlying foundational dynamics. In the second stage, the model is fine-tuned on a smaller, high-quality dataset to learn the mean velocity. In this stage, we adopt the strategy from (Geng et al., 2025), where we blend the instantaneous and average fields by randomly

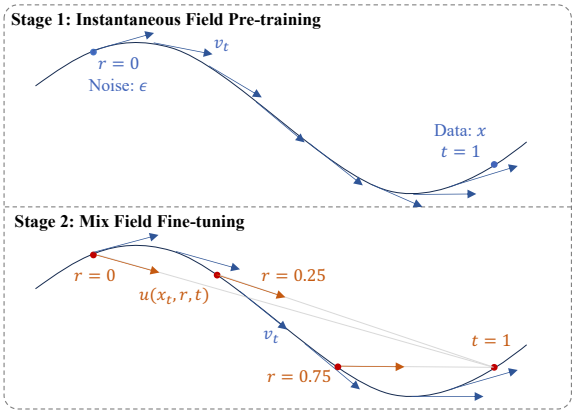

Figure 3: Illustration of the underlying instantaneous field v_t, which models tangents of the flow trajectory, and the average velocity field $u(x_t, r, t)$, which captures long displacements. Our training curriculum encourages the model to first learn the foundational instantaneous dynamics, then gradually adapt to mean flows for fast and faithful generation.

setting $r = t$. As illustrated in Eq. 2, this operation degenerates the MF objective into standard flow matching. As such, the network can provide an effective derivative approximation during fine-tuning by leveraging the knowledge acquired in pre-training. Furthermore, by combining two flows, the model can stably adapt to the average field for fast few-step generation, while preserving the multi-step synthesis performance through adherence to instantaneous velocity. We will demonstrate the effectiveness of our training curriculum in the experimental section.

4 Experiments

4.1 Datasets

We collect a large-scale audio-text dataset to train MeanAudio, including AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2024), AudioSet (Gemmeke et al., 2017), VGGSound (Chen et al., 2020), and LP-MusicCaps (MC & MTT parts) (Doh et al., 2023). Among them, only AudioCaps provides high-quality human-annotated textual captions, while others are weakly labeled using Large Language Models (Schulman et al., 2022) or Multimodal Large Language Models (Chu et al., 2024). During training, we truncate all audio into 10 seconds. For audios longer than 10s, we crop at most 5 non-overlapping 10s segments. In total, we collect approximately 2.8M audio-text pairs for training, with a total duration of 8k hours. A detailed dataset description is given in Appendix D.

We use the test split of AudioCaps as the eval-

Model	Prams*	NFE	FAD ↓	FD ↓	KL ↓	IS ↑	CLAP ↑	RTF† ↓
Diffusion- and Flow-based TTA Models								
AudioLDM-L-Full (Liu et al., 2023a)‡	739M	200	4.32	29.50	1.68	8.17	0.208	2.935
Tango-Full-FT (Ghosal et al., 2023)‡	866M	200	2.68	15.64	1.24	8.78	0.291	2.382
EzAudio-XL (Hai et al., 2025)‡	875M	200	3.64	14.98	1.29	11.38	0.314	1.718
Stable Audio Open (Evans et al., 2025)‡	1050M	200	4.19	39.14	2.36	10.07	0.209	2.867
TangoFlux (Hung et al., 2024)‡	516M	50	2.41	20.65	1.27	<u>12.81</u>	<u>0.318</u>	<u>0.449</u>
IMPACT-Large (Huang et al., 2025)**	427M	64	**1.17**	14.72	**1.07**	10.53	-	-
ETTA (Lee et al., 2024)**	1440M	100	-	**13.12**	1.42	**14.36**	-	-
GenAU-Large-Full (Haji-Ali et al., 2024)‡	1250M	200	<u>2.07</u>	14.58	1.36	10.43	0.300	1.612
Accelerated TTA Models								
AudioLCM (Liu et al., 2024b)‡	160M	*1*	*4.70*	*24.59*	*1.67*	*8.04*	*0.200*	*0.023*
		2	2.16	19.81	1.46	10.05	0.245	0.026
		5	1.76	20.12	1.46	10.36	0.256	0.029
ConsistencyTTA (Bai et al., 2023)◇,‡	559M	*1*	*<u>2.31</u>*	*<u>22.16</u>*	*<u>1.44</u>*	*<u>9.15</u>*	*<u>0.268</u>*	*<u>0.017</u>*
		2	2.68	22.25	1.42	9.20	0.270	0.022
		5	3.38	24.60	1.49	9.12	0.302	0.038
AudioTurbo (Zhao et al., 2025)**	1100M	5	-	22.18	1.29	9.40	0.298	-
MeanAudio (Ours)◇	120M	*1*	***1.77***	***14.30***	***1.32***	***10.02***	***0.290***	***0.013***
		2	1.84	13.57	1.27	10.70	0.305	0.015
		5	2.12	13.43	1.25	11.18	0.314	0.024
		25	2.30	<u>13.38</u>	<u>1.25</u>	11.23	**0.319**	**0.083**

Table 1: Objective evaluation results on AudioCaps test set. The single-step generation results are in *italic*. The best multi-step and single-step generation results are **bolded**, the second-best are <u>underlined</u>. ∗: The parameter count refers to the diffusion backbone, excluding text encoders, VAE, and vocoders. †: The real-time factor (RTF) is evaluated over 100 generations samples on an NVIDIA RTX 3090. ∗∗: Closed-source models, results are transcribed from the original paper. ‡: Open-source models, results are re-evaluated using the official checkpoint. ◇: These models integrated CFG into training, resulting in *True* NFEs, for other models, NFE should actually be doubled.

uation set, which contains 957 audio clips. Each audio of the test set is paired with 5 textual captions, and we randomly select 1 caption for audio generation. To further evaluate MeanAudio's music generation capability, we test its performance on the test set of MusicCaps (Agostinelli et al., 2023), with results reported in Appendix C.3.

Models	Size	NFE	OVL↑	REL↑
MeanAudio (Ours)	120M	1	**4.03**$\pm$0.72	**4.29**$\pm$0.71
ConsistencyTTA	559M	1	3.32$\pm$0.95	3.41$\pm$1.03
AudioLCM	160M	1	3.47$\pm$0.89	3.31$\pm$1.13
MeanAudio (Ours)	120M	25	**4.19**$\pm$0.70	**4.44**$\pm$0.66
TangoFlux	516M	50	4.02$\pm$0.87	4.22$\pm$0.85
AudioLDM-L-Full	739M	200	3.25$\pm$1.00	3.33$\pm$1.08
GenAU-Large-Full	1250M	200	3.66$\pm$0.90	3.79$\pm$0.94

Table 2: Subjective evaluation results

4.2 Implementation Details

MeanAudio has $N_1 = 4$ multi-modal blocks and $N_2 = 8$ single-modal blocks. The hidden dimension of the transformer is set to 448, and the network has a total of 120M parameters. For integrated classifier-free guidance, We set $\omega = 0.3$ and $\kappa = 0.9$, resulting in an effective guidance scale of $\frac{\omega}{1-\kappa} = 3$. We sample timesteps (t, r) according to a logit-normal distribution (Esser et al., 2024), with $\mu = 0.4$ and $\sigma = 1$. Given a sampled pair, we assign the larger value to t and the smaller value to r. Additionally, 75% of the samples are randomly set with $r = t$ to mix the vanilla instantaneous flow field with the mean flow field.

Following the curriculum described in Section 3.5, the model is first pre-trained on the full dataset for 400k iterations with the standard flow matching objective, and subsequently fine-tuned on AudioCaps for 200k iterations using the mixed flow objective. During both stages, we use a learning rate of 1e-4 with a linear warm-up of 1,000 steps. A step decay schedule reduces the learning rate to 10% of its original value at 80% and 90% of total steps. Batch sizes are set to 256 and 32 for pre-training and fine-tuning, respectively. All experiments are conducted on four NVIDIA RTX 3090 GPUs, requiring approximately 48 hours for pre-training and 22 hours for fine-tuning.

For subjective evaluation, we adopt standard TTA evaluation metrics: Fréchet Distance (FD), Fréchet Audio Distance (FAD), Kullback–Leibler Divergence (KL), Inception Score (IS), and CLAP score. For objective evaluation, we recruit ten au-

dio professionals to carry out a rating process, following (Liu et al., 2023a, 2024a). Specifically, the generated samples are rated based on overall quality (OVL) and relevance to the input text (REL) on a scale of 1 to 5. To further evaluate the model's inference speed, we also report the Real-time Factor (RTF) of the system, which denotes the ratio between the total time a system takes to synthesize an audio and the duration of the audio. More details about evaluation can be found in Appendix B.

4.3 Main Results

We compare the performance of MeanAudio with other TTA models under both few-step (< 10 NFEs) and multi-step (≥ 10 NFEs) generation. For few-step generation, we compare with SOTA accelerated TTA models. For multi-step generation, we compare with best-performing diffusion and flow-based models.

As shown in Table 1, for objective metrics, MeanAudio demonstrates SOTA performance under single-step generation, achieving an FD of 14.30, a KL of 1.32, an IS of 10.02, and a CLAP score of 0.290, outperforming all accelerated models such as ConsistencyTTA and AudioLCM by large margins. Remarkably, its single-step generation performance even surpasses some multi-step baselines that require hundreds of function evaluations. Moreover, MeanAudio achieves the fastest real-time factor of 0.013, representing a 100x speedup over the SOTA open-sourced diffusion-based model, GenAU, which has an RTF of 1.612.

As the number of sampling steps increases, MeanAudio's generation quality also improves, revealing a trade-off between inference speed and output fidelity. Its two-step and five-step generations consistently outperform previous SOTA methods, highlighting its strong few-step synthesis capabilities. For multi-step generation, MeanAudio remains competitive with the best-performing audio systems. With only 25 synthesis steps, it achieves an FD of 13.38, a KL of 1.25, and a CLAP score of 0.318. It is worth noting that MeanAudio contains only 120M parameters, whereas other TTA systems typically exceed 500M.

Furthermore, the subjective evaluation results in Table 2 align well with these objective findings. For single-step generation, MeanAudio achieves the highest human-rated overall quality and relevance, obtaining OVL and REL scores of 4.03 and 4.29, respectively, surpassing ConsistencyTTA and AudioLCM. Under the multi-step settings, MeanAudio attains an OVL of 4.19 and a REL of 4.44, outperforming all diffusion-based competitors. These evaluation results confirm that MeanAudio is capable of generating high-quality and semantically coherent audio, while being both fast and lightweight.

Models	NFE	FD↓	KL↓	IS↑	CLAP↑
FluxAudio (Ours)	1	99.10	4.48	2.45	-0.007
MeanAudio-Scratch		16.13	1.36	9.62	0.285
w. Pre-training		**14.30**	**1.32**	**10.02**	**0.290**
FluxAudio (Ours)	25	15.70	1.30	**11.24**	**0.328**
MeanAudio-Scratch		14.41	1.32	10.79	0.314
w. Pre-training		**13.38**	**1.25**	11.23	0.318

Table 3: Ablation study of the training curriculum.

Ratio of $r = t$	NFE	FD↓	KL↓	IS↑	CLAP↑
0%	1	138.34	5.91	1.19	-0.061
25%		23.73	1.75	7.30	0.213
50%		16.18	1.39	9.45	**0.287**
75%		**16.13**	**1.36**	**9.62**	0.285
0%	25	150.75	6.28	1.14	-0.050
25%		22.15	1.63	7.80	0.241
50%		15.42	1.37	9.78	0.296
75%		**14.41**	**1.32**	**10.79**	**0.314**

Table 4: Ablation study of the flow mix-up ratio.

4.4 Ablation Studies

We conduct a comprehensive ablation study covering training strategies, architectural designs, and flow configurations to identify best practices for building a MeanFlow-based audio generator. Here, all experiments except the training curriculum are trained from scratch on AudioCaps for 200k steps.

Instantaneous-to-Mean Curriculum. We begin by studying the benefits of the proposed training strategy. For this, we first evaluate the performance of the model trained with the standard flow matching objective, denoted as FluxAudio. As shown in Table 3, FluxAudio delivers strong performance in multi-step generation but performs poorly under single-step scenarios. For instance, it achieves a CLAP score of 0.328 with 25 NFEs, while its one-step synthesis yields only -0.007. This large performance gap arises because the vanilla flow matching objective guides only toward the instantaneous velocity field and small displacements. Secondly, we trained MeanAudio using the mixed flow objective as described in Section 3.5, but without initializing it from FluxAudio, which we denote as MeanAudio-Scratch. As shown in Table 3, training MeanAudio with a mixed flow field significantly enhances its single-step generation, as the FD and IS improve to 16.13 and 9.62 at NFE =

1. Finally, training MeanAudio on top of FluxAudio with mixed flow fields further improves its performance in both single- and multi-step generation. As illustrated in Table 3, at NFE = 1, the FD, CLAP score, and IS increase to 14.30, 10.02, and 0.290, respectively. Additionally, at NFE = 25, these scores rise to 13.38, 11.23, and 0.318, which are comparable to FluxAudio. These results demonstrate that our proposed instantaneous-to-mean curriculum can effectively enhance generation quality under both single- and multi-step scenarios.

Flow Field Mix-up. We then investigated the effectiveness of the flow field mix-up by changing the ratio of $r = t$. Remember that this equals the percentage of using the vanilla flow matching objective during the fine-tuning stage. As illustrated in Table 4, when the flow ratio $r = t$ is set to 0 (using only the mean flow objective), training becomes highly unstable and fails to converge. Setting the ratio to 25% stabilizes training, but we observe only a modest improvement in multi-step generation performance, with FD increasing by just 2.5%. As the flow ratio further increases, convergence accelerates, and generation performance also improves in both single- and multi-step settings. Specifically, when the ratio increases from 25% to 75%, the single-step CLAP score increases from 0.213 to 0.285, and the multi-step IS improves from 7.80 to 10.79. These results suggest that in audio latent modeling, jointly learning the instantaneous and mean velocity can stabilize training and enhance performance in both single-step and multi-step generation, as the two flows complement each other.

Architectural Designs. We then study the effectiveness of different model components. As shown in Table 5, removing the CLAP encoder causes a clear drop in text–audio alignment, as the CLAP score drops from 0.285 to 0.270 and KL degrades from 1.36 to 1.42. These results underscore the importance of the dual-encoder design for semantic coherence. Similarly, excluding RoPE also reduces the CLAP score from 0.285 to 0.276, indicating that rotary positional embeddings can help preserve relative temporal order and strengthen alignment. Finally, removing ConvMLP degrades both the generation fidelity and diversity, as the FD score drops from 16.13 to 17.22 and IS worsens from 9.62 to 9.21, which highlights its role in temporal modeling and local feature extraction. We also include a subjective evaluation of the dual-encoder design and experiments on the scalability of MeanAudio, which can be found in Appendix C.

Models	NFE	FD↓	KL↓	IS↑	CLAP↑
MeanAudio (Ours)	1	16.13	**1.36**	9.62	**0.285**
w/o CLAP		16.21	1.42	9.80	0.270
w/o RoPE		**15.87**	1.41	**9.83**	0.276
w/o ConvMLP		17.22	1.38	9.21	0.283

Table 5: Ablation study of the network design.

CFG Scale	NFE	FD↓	KL↓	IS↑	CLAP↑
1 (No CFG)	1	28.02	2.03	4.90	0.185
2		16.33	1.44	8.65	0.270
3		**16.13**	**1.36**	9.62	**0.285**
4		16.23	**1.36**	9.52	0.284
5		17.09	1.40	**9.78**	0.284

Table 6: Ablation study of the CFG scale.

Integrated Classifier-Free Guidance. We finally conducted an ablation study across different guidance scales to evaluate the effectiveness of CFG. As shown in Table 6, increasing the CFG scale from 1 (no guidance) to 3 leads to substantial improvements in both generation quality and semantic alignment, where FD, IS, and CLAP score have increased by 42.4%, 96.3%, and 54.1%, respectively. Further increasing the guidance scale to 4 and 5 yields marginal improvements in IS, but slightly degrades FD, KL, and CLAP, suggesting a trade-off between generation diversity and prompt adherence. These results suggest that a moderate guidance scale around 3 is the optimal configuration for single-step audio generation. Note that the CFG is integrated into MeanAudio's training and incurs no additional cost during sampling.

5 Conclusion

In this work, we present MeanAudio, a novel MeanFlow-based fast and faithful text-to-audio generator. Built upon an enhanced Flux-style latent transformer, MeanAudio regresses the guided average velocity field during training, enabling fast generation by directly mapping from the start to the endpoint of the flow trajectory. To enhance training stability and generation quality, MeanAudio adopts an instantaneous-to-mean curriculum with flow-field mix-up, facilitating stable optimization and rapid convergence. Extensive experiments demonstrate that MeanAudio delivers SOTA performance in single-step generation and competitive results in multi-step synthesis. Comprehensive ablation studies further highlight the importance of architectural design, training curriculum, and flow configurations, providing valuable insights for developing faster and stronger audio systems.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. U23B2018), Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102 and Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG1100).

Limitations

While MeanAudio achieves strong performance in both single-step and multi-step generation, several limitations remain. Firstly, the training data used for MeanAudio are cropped to 10-second segments, which may limit its ability to generate variable-length or long-form audio. Secondly, since MeanAudio is trained exclusively on public datasets, which can be noisy to some extent, its generation quality may still lag behind models trained on larger and cleaner proprietary datasets. Thirdly, MeanAudio currently focuses on text-to-sound and music generation, and therefore lacks the ability to produce intelligible speech or handle fine-grained linguistic content.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Michael S Albergo and Eric Vanden-Eijnden. 2023. Building normalizing flows with stochastic interpolants. *Proc. ICLR*.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2025. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE Transactions on Audio, Speech and Language Processing*.

Yatong Bai, Trung Dang, Dung Tran, Kazuhito Koishida, and Somayeh Sojoudi. 2023. ConsistencyTTA: Accelerating diffusion-based text-to-audio generation with consistency distillation. *Proc. Interspeech*.

BlackForestLabs. 2024. Flux. https://github.com/black-forest-labs/flux.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *Proc. ICASSP*.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. 2024. Taming multimodal joint training for high-quality video-to-audio synthesis. *Proc. CVPR*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. LP-MusicCaps: Llm-based pseudo music captioning. *Proc. ISMIR*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Proc. ICML*.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable audio open. *Proc. ICASSP*.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. 2024. Flux that plays music. *arXiv preprint arXiv:2409.00587*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. *Proc. ICASSP*.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. 2025. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-audio generation using instruction guided latent diffusion model. *Proc. ACM MM*.

Wenhao Guan, Kaidi Wang, Wangjin Zhou, Yang Wang, Feng Deng, Hui Wang, Lin Li, Qingyang Hong, and Yong Qin. 2024. LAFMA: A latent flow matching model for text-to-audio generation. *Proc. Interspeech*.

Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, and Dong Yu. 2025. EzAudio: Enhancing text-to-audio generation with efficient diffusion transformer. *Proc. Interspeech*.

Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, and Vicente Ordonez. 2024. Taming data and transformers for audio generation. *arXiv preprint arXiv:2406.19388*.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and 1 others. 2017. Cnn architectures for large-scale audio classification. *Proc. ICASSP*.

Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. The benefit of temporally-strong labels in audio event classification. *Proc. ICASSP*.

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

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

Kuan-Po Huang, Shu-wen Yang, Huy Phan, Bo-Ru Lu, Byeonggeun Kim, Sashank Macha, Qingming Tang, Shalini Ghosh, Hung-yi Lee, Chieh-Chi Kao, and 1 others. 2025. Impact: Iterative mask-based parallel decoding for text-to-audio generation with diffusion modeling. *Proc. ICML*.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023b. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *Proc. ICML*.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. 2024. TangoFlux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. *Proc. NAACL*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *Proc. ICLR*.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28.

Zhifeng Kong, Sang-gil Lee, Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, Rafael Valle, Soujanya Poria, and Bryan Catanzaro. 2024. Improving text-to-audio models with synthetic captions. *Proc. Interspeech*.

Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. 2009. Evaluation of algorithms using games: The case of music tagging. In *Proc. ISMIR*.

Sang-gil Lee, Zhifeng Kong, Arushi Goel, Sungwon Kim, Rafael Valle, and Bryan Catanzaro. 2024. ETTA: Elucidating the design space of text-to-audio models. *Proc. ICML*.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. BigVGAN: A universal neural vocoder with large-scale training. *Proc. ICLR*.

Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li. 2024. BATON: Aligning text-to-audio model with human preference feedback. *Proc. IJCAI*.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling. *Proc. ICLR*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023a. AudioLDM: Text-to-audio generation with latent diffusion models. *Proc. ICML*.

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

Huadai Liu, Rongjie Huang, Yang Liu, Hengyuan Cao, Jialei Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. 2024b. AudioLCM: Text-to-audio generation with latent consistency models. *Proc. ACM MM*.

Huadai Liu, Jialei Wang, Rongjie Huang, Yang Liu, Heng Lu, Zhou Zhao, and Wei Xue. 2025. FlashAudio: Rectified flows for fast and high-fidelity Text-to-Audio Generation. *Proc. ACL*.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2023b. Flow straight and fast: Learning to generate and transfer data with rectified flow. *Proc. ICLR*.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *Proc. ACM MM*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. WavCaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zachary Novack, Zach Evans, Zack Zukowski, Josiah Taylor, CJ Carr, Julian Parker, Adnan Al-Sinan, Gian Marco Iodice, Julian McAuley, Taylor Berg-Kirkpatrick, and 1 others. 2025a. Fast text-to-audio generation with adversarial post-training. *arXiv preprint arXiv:2505.08175*.

Zachary Novack, Ge Zhu, Jonah Casebeer, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J Bryan. 2025b. Presto! distilling steps and layers for accelerating music generation. *Proc. ICLR*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. *Proc. ICCV*.

Yansong Peng, Kai Zhu, Yu Liu, Pingyu Wu, Hebei Li, Xiaoyan Sun, and Feng Wu. 2025. Flow-anchored consistency models. *arXiv preprint arXiv:2507.03738*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Proc. CVPR*.

Koichi Saito, Dongjun Kim, Takashi Shibuya, Chieh-Hsin Lai, Zhi Zhong, Yuhta Takida, and Yuki Mitsufuji. 2025. SoundCTM: Unifying score-based and consistency models for full-band text-to-sound generation. *Proc. ICLR*.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, and 1 others. 2022. Introducing ChatGPT. *OpenAI Blog*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency Models. *Proc. ICML*.

Stanley Smith Stevens, John Volkmann, and Edwin B Newman. 1937. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*.

Rafael Valle, Rohan Badlani, Zhifeng Kong, Sang-gil Lee, Arushi Goel, Sungwon Kim, Joao Felipe Santos, Shuqi Dai, Siddharth Gururani, Aya Aljafari, and 1 others. 2025. Fugatto 1: Foundational generative audio transformer opus 1. In *Proc. ICLR*.

Fu-Yun Wang, Ling Yang, Zhaoyang Huang, Mengdi Wang, and Hongsheng Li. 2024. Rectified diffusion: Straightness is not your need in rectified flow. *Proc. ICLR*.

Zehan Wang, Ke Lei, Chen Zhu, Jiawei Huang, Sashuai Zhou, Luping Liu, Xize Cheng, Shengpeng Ji, Zhenhui Ye, Tao Jin, and 1 others. 2025. T2a-feedback: Improving basic capabilities of text-to-audio generation via fine-grained ai feedback. *Proc. ACL*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *Proc. ICASSP*.

Mu Yang, Bowen Shi, Matthew Le, Wei-Ning Hsu, and Andros Tjandra. 2024. Audiobox tta-rag: Improving zero-shot and few-shot text-to-audio with retrieval-augmented generation. *Proc. Interspeech*.

Yi Yuan, Dongya Jia, Xiaobin Zhuang, Yuanzhe Chen, Zhengxi Liu, Zhuo Chen, Yuping Wang, Yuxuan Wang, Xubo Liu, Xiyuan Kang, and 1 others. 2025. Sound-VECaps: Improving audio generation with visual enhanced captions. *Proc. ICASSP*.

Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. 2024. Retrieval-augmented text-to-audio generation. *Proc. ICASSP*.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Proc. NeurIPS*.

Junqi Zhao, Jinzheng Zhao, Haohe Liu, Yun Chen, Lu Han, Xubo Liu, Mark Plumbley, and Wenwu Wang. 2025. AudioTurbo: Fast text-to-audio generation with rectified diffusion. *Proc. Interspeech*.

A Related Work

A.1 Text to Audio Generation

Text-to-Audio Generation (TTA) focuses on generating sound based on text inputs. Current TTA models are often based on the Latent Diffusion Model (LDM) (Rombach et al., 2022) architecture. Among all, AudioLDM (Liu et al., 2023a) pioneered this by training a U-Net to perform denoising conditioned on CLAP (Wu et al., 2023) embeddings. Tango (Ghosal et al., 2023) further improves the instruction following abilities by using the Large Language Model (LLM) FLAN-T5 (Chung et al., 2024) as the text encoder. More recently, Stable-Audio-Open (Evans et al., 2025) adopted the Diffusion Transformer (DiT) (Peebles and Xie, 2023) to generate variable-length, full-band audio samples. Meanwhile, LAFMA (Guan et al., 2024) and TangoFlux (Hung et al., 2024) integrated Flow Matching (Liu et al., 2023b) to improve the efficiency and fidelity of audio generation. Additionally, recent work has explored Reinforcement Learning (RL) (Liao et al., 2024; Majumder et al., 2024; Wang et al., 2025) and Retrieval-Augmented Generation (RAG) (Yuan et al., 2024; Yang et al., 2024) to enhance generation controllability and output quality. While these models achieve high fidelity and controllability, they often suffer from slow inference due to the iterative sampling process inherent to diffusion- and flow-based models.

A.2 TTA with Inference Acceleration

To accelerate TTA inference, ConsistencyTTA first integrates Consistency Distillation (Song et al., 2023) on TTA Latent Diffusion Models, where a teacher model, Tango (Majumder et al., 2024), provides supervision to train a distilled few-step student generator. Subsequently, AudioLCM (Liu et al., 2024b) enriches this framework by employing a multi-step Ordinary Differential Equation (ODE) solver, while SoundCTM (Saito et al., 2025) introduces a novel feature distance to enable flexible single-step and multi-step generation. In parallel, Presto (Novack et al., 2025b) proposes a dual-faceted distillation strategy that reduces both the number of sampling steps and model parameters to improve inference efficiency. FlashAudio (Liu et al., 2025) and AudioTurbo (Zhao et al., 2025) explored the use of Rectified Flow (Liu et al., 2023b) and Rectified Diffusion (Wang et al., 2024) to learn straight generative paths for rapid audio synthesis. Meanwhile, Stable-Audio-Small (Novack et al., 2025a) employs contrastive post-training with adversarial loss to construct a compact few-step generator. In this work, we investigate the use of MeanFlow to develop a TTA model that achieves strong performance in both single-step and multi-step generation.

Despite encouraging results, these distillation-based approaches often require significant computational resources, as online methods require holding 2-3 full models in memory at the same time, and offline methods rely on large-scale generation and storage of teacher trajectories before training. Moreover, their performance is inevitably limited by the pre-trained teacher model and the consistency constraint do not provide properties of the underlying ground-truth field that should guide learning, which could lead to unstable training (Geng et al., 2025).

B Evaluation Details

In this section, we first provide a detailed explanation of the objective evaluation metrics, followed by a description of our human evaluation process.

Among all metrics, FD and FAD measure the distance between the generated audio distribution and the real audio distribution. A low FD indicates that the generated audio is realistic and closely resembles the reference audio. KL evaluates how semantically similar the generated audio is to the reference audio. IS measures the diversity and quality of the generated samples, and CLAP score[1] measures how the generated audio align with the textual prompt. FD, IS, and KL are calculated based on the state-of-the-art audio tagger PANNs (Kong et al., 2020), while FAD is calculated by VGGish (Hershey et al., 2017).

For human subjective evaluation, we selected 10 audio experts to perform the listening test. Each participant is presented with 10 audio samples generated by different models along with their corresponding prompts. The captions are randomly selected from the AudioCaps test set, and the order of the audio samples is also randomized. Participants are asked to provide OVL and REL scores for each sample. An illustration of the evaluation platform is shown in Figure 4.

.

[1]The CLAP score is calculated based on the checkpoint: https://huggingface.co/lukewys/laion_clap/blob/main/music_speech_audioset_epoch_15_esc_89.98.pt

Model	Prams	NFE	FAD ↓	FD ↓	KL ↓	IS ↑	CLAP ↑
MeanAudio-S-AC	120M		**1.77**	14.30	1.32	10.02	0.290
MeanAudio-S-Full	120M	1	2.85	**14.19**	1.27	10.69	0.314
MeanAudio-L-Full	480M		2.17	14.22	**1.23**	**11.24**	**0.316**
MeanAudio-S-AC	120M		**2.30**	**13.38**	1.25	11.23	0.319
MeanAudio-S-Full	120M	25	3.51	**13.38**	1.22	12.11	0.334
MeanAudio-L-Full	480M		2.59	14.06	**1.21**	**12.46**	**0.339**

Table 7: Experiments on Data & Model Scaling. Results are evaluated on AudioCaps test set.

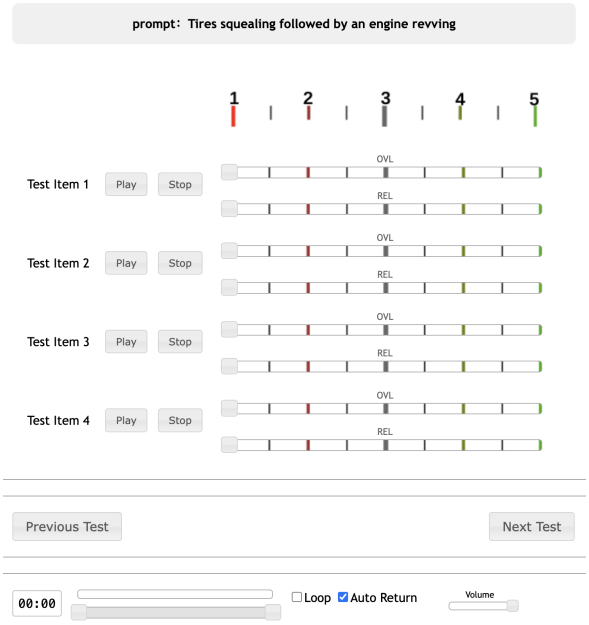

Figure 4: Screenshot of the subjective evaluation platform.

C Addtional Experiments

In this section, we present several additional experiments, including a detailed human evaluation of the dual-encoder architecture, an analysis of the scalability of MeanAudio, and an assessment of its music generation capability.

C.1 Human Evaluation on Dual Encoders

To more comprehensively evaluate and understand the benefits of the dual-encoder architecture, we designed a more detailed human evaluation protocol. We first constructed a refined evaluation subset, aiming to better reflect the model's ability to follow long and complex textual descriptions. Specifically, we selected captions from the AudioCaps test set that contain more than 20 words, forming a subset named AudioCaps-Long. We observed that such captions typically describe at least four distinct sound events, posing a greater challenge for models in terms of instruction comprehension and multi-event generation. From AudioCaps-Long, we then randomly sampled 10 captions and asked 3 human listeners to evaluate the relevance (REL) of the generated audio outputs. During this evaluation, the listeners were instructed to pay particular attention to fine-grained sound events and their temporal order when assigning scores.

Models	Size	NFE	Encoder	REL↑
MeanAudio	120M	1	T5+CLAP	**4.10**$\pm$0.37
MeanAudio	120M	1	T5	3.75$\pm$0.51

Table 8: Subjective evaluation on dual encoders

As shown in Table 8, using both FLAN-T5 and CLAP consistently yields better performance than employing T5 alone. The dual-encoder model tends to capture subtle sound events more accurately, such as a dog barking at the beginning or a child crying toward the end of the audio, confirming the effectiveness of our architectural design.

C.2 Experiments on Scalability

We further investigate the scalability of MeanAudio with respect to both data and model size. For data scaling, we initialize MeanAudio from FluxAudio and train it on the entire 2.8M-sample dataset using the mixed flow objective for 400k iterations with a batch size of 256. The resulting large-scale variant is denoted as MeanAudio-S-Full. For model scaling, we increase the hidden dimension to 896, resulting in a model with 448M parameters, referred to as MeanAudio-L-Full. We compare these two new variants with MeanAudio-S-AC, which was initialized from FluxAudio and was trained on AudioCaps with the mixed-flow objective.

As shown in Table 7, both data scaling and model scaling substantially improve MeanAudio's generation quality. When trained on the full 2.8M-sample dataset, the model (MeanAudio-S-Full) achieves clear gains over MeanAudio-S-AC under

Model	Prams*	NFE	FAD ↓	FD ↓	KL ↓	IS ↑	CLAP ↑
Accelerated TTA Models							
AudioLCM (Liu et al., 2024b)†	160M	1	5.98	40.68	1.62	2.33	0.149
ConsistencyTTA (Bai et al., 2023)†	559M	1	4.89	67.30	1.87	1.69	0.138
MeanAudio-S-AC (Ours)[1]	120M	1	3.38	27.96	1.71	2.29	0.200
MeanAudio-S-Full (Ours)[2]	120M	1	**1.26**	**11.82**	**1.11**	**2.90**	**0.302**
Diffusion- and Flow-based TTA Models							
FluxMusic (Fei et al., 2024)‡	2100M	200	1.43	-	1.25	2.98	-
AudioLDM-2-Large (Liu et al., 2024a)‡	712M	200	2.93	16.34	1.40	2.59	-
Stable-Audio-Open (Evans et al., 2025)‡	1050M	200	3.51	36.42	1.56	2.93	-
ETTA (Lee et al., 2024)‡	1440M	100	1.91	10.06	**1.04**	**3.32**	-
MeanAudio-S-AC (Ours)	120M	25	3.05	24.73	1.71	2.56	0.221
MeanAudio-S-Full (Ours)	120M	25	**1.21**	**10.05**	1.08	3.01	**0.315**

Table 9: Objective evaluation results on MusicCaps test set. †: Results are evaluated using the officially released checkpoint. ‡: Results are transcribed from the original paper. [1]: MeanAudio-S-AC: Model initialized from FluxAudio and fine-tuned with the mixed flow objective on the training set of AudioCaps. [2]: Model initialized from FluxAudio and fine-tuned with the mixed flow objective on the full training set. Detailed explanation about these models found in Appendix C.2.

both single-step and multi-step settings. At NFE = 1, the IS and CLAP scores increase from 10.02 to 10.69 and 0.290 to 0.314, respectively, indicating stronger prompt adherence and improved synthesis fidelity. Further scaling the model size to 480M parameters (MeanAudio-L-Full) yields consistent improvements, with IS = 11.24 and CLAP = 0.316, surpassing all other single-step audio generators and approaching the performance of multi-step systems. A similar trend is observed in the multi-step scenario, where scaling enhances both fidelity and semantic alignment: MeanAudio-L-Full achieves CLAP = 0.339 and IS = 12.46, the best among all configurations. These results demonstrate that MeanAudio can benefit from both data and model scaling, confirming its scalability and capacity to leverage larger datasets and architectures for higher-quality audio generation.

C.3 Text-to-Music Generation

Finally, we evaluate MeanAudio's music generation abilities, where we assess its performance on the test split of MusicCaps (Agostinelli et al., 2023). As summarized in Table 9, MeanAudio significantly outperforms existing accelerated text-to-audio systems while achieving performance on par with or superior to larger diffusion- and flow-based models.

Under single-step generation, MeanAudio-S-AC achieves an FD of 27.96 and an IS of 2.29, significantly surpassing AudioLCM and ConsistencyTTA by large margins. By further scaling the training data, MeanAudio continues to exhibit substantial improvements in music generation performance. Specifically, MeanAudio-S-Full achieves an FD of 11.82 and an IS of 2.90, showcasing its enhanced capability to synthesize high-quality music within a single generation step.

When the sampling steps are increased to 25, MeanAudio-S-Full further improves to a FAD of 1.21, FD of 10.05, and CLAP of 0.315, reaching performance comparable to state-of-the-art diffusion-based music generation systems that typically require hundreds of iterations. These results highlight the strong generative capability of MeanAudio and demonstrate the effectiveness of the MeanFlow framework in producing perceptually rich music signals.

C.4 Experiments on Transformer Backbone

We further investigate the impact of the transformer backbone by replacing the Flux-style MMDiT architecture with a standard DiT design.

Specifically, we remove all MMDiT layers and substitute them with standard DiT layers. To maintain a comparable parameter budget, we increase the network depth from 12 to 16 layers. For the standard DiT variant, the text features are processed as follows: (1) mean pooling is applied over the sequence dimension of the FLAN-T5 features $y_{\text{T5}} \in \mathbb{R}^{N \times d_{\text{T5}}}$ to obtain $\bar{y}_{\text{T5}} \in \mathbb{R}^{1 \times d_{\text{T5}}}$; (2) the pooled feature is concatenated with the CLAP text embedding $y_{\text{CLAP}} \in \mathbb{R}^{1 \times d_{\text{CLAP}}}$, yielding $y_{\text{T5+CLAP}} \in \mathbb{R}^{1 \times (d_{\text{T5}} + d_{\text{CLAP}})}$; (3) the resulting

Model	Params	NFE	FAD ↓	FD ↓	KL ↓	IS ↑	LAION-CLAP ↑
MeanAudio-DiT (Standard DiT)	136M	1	2.31	24.22	1.99	6.87	0.213
MeanAudio (Flux-style MMDiT)	120M	1	**1.77**	**16.13**	**1.36**	**9.62**	**0.285**

Table 10: Comparison between the standard DiT backbone and the proposed Flux-style MMDiT.

feature is projected and injected into the DiT backbone via AdaLN.

We denote this variant as **MeanAudio-DiT**. The training setup is kept identical to the main model (MeanAudio), using AudioCaps for 200k steps with a batch size of 32. The results show that replacing the Flux-style MMDiT with a standard DiT backbone leads to consistent degradation across all evaluation metrics. In particular, MeanAudio-DiT exhibits higher FAD, FD, and KL, as well as lower IS and CLAP scores, indicating inferior audio quality and weaker text-audio alignment. These findings suggest that the Flux-style MMDiT architecture plays a critical role in improving both acoustic fidelity and semantic consistency. We attribute this improvement to its dual-stream design, which enables more effective refinement of text representations, and its joint attention mechanism, which facilitates richer cross-modal interactions.

D Dataset Details

In this section, we present a detailed description of the datasets used in the training of MeanAudio. Table 11 provides an overall statistics of all datasets.

AudioCaps. AudioCaps (Kim et al., 2019) contains approximately 50k audio-text pairs, where each audio contains one human-labelled high-quality caption.

WavCaps. WavCaps (Mei et al., 2024) comprises 400k audio samples collected from multiple sources, including BBC Sound Effects, [2], FreeSound [3] , SoundBible [4] and AudioSet-Strong (Hershey et al., 2021). Each audio has one caption generated by ChatGPT.

AudioSet. AudioSet (Gemmeke et al., 2017) contains approximately 2M audio samples. However, these audio segments only contain ground-truth labels. We use the caption from AudioSetCaps (Bai et al., 2025) to train our TTA models.

VGGSound. VGGSound (Chen et al., 2020) contains about 200k audio samples. Like AudioSet, the audio is only paired with labels. We directly use labels to train the TTA models.

LP-MusicCaps. LP-MusicCaps (Doh et al., 2023) is a large-scale weakly-labeled music caption dataset, sourced from MusicCaps (MC) (Agostinelli et al., 2023), MagnaTagATune (MTT) (Law et al., 2009), and Million-Song-Dataset (MSD). We use the MC and MTT splits of LP-MusicCaps, containing a total of 25k music segments.

[2] https://sound-effects.bbcrewind.co.uk
[3] https://freesound.org
[4] https://soundbible.com

Dataset Name	# Clips	Caption Source	Duration
AudioCaps (Kim et al., 2019)	49k	Human (Kim et al., 2019)	130h
AudioSet (Gemmeke et al., 2017)	1.7M	MLLM† (Bai et al., 2025)	4722h
FreeSound	612k	LLM‡ (Mei et al., 2024)	1700h
BBC-Sound-Effects	121k	LLM (Mei et al., 2024)	336h
AudioSet-Strong (Hershey et al., 2021)	106k	LLM (Mei et al., 2024)	294h
VGGSound (Chen et al., 2020)	178k	Human (Chen et al., 2020)	494h
LP-MusicCaps-MC	2.3k	LLM (Doh et al., 2023)	6.3h
LP-MusicCaps-MTT	43k	LLM (Doh et al., 2023)	119h
Total	2.8M	-	7801h

Table 11: Dataset Statistics. † MLLM: Multimodal Large Language Models. ‡ LLM: Large Language Models.