A Resource-Constrained Text-to-Audio Generation Model Using Linear Diffusion Transformer and Locally Optimized Mamba

Biao Yang, *Senior Member, IEEE*, Ruiyang Shen, Yang Chen, Rongrong Ni, and Wenwu Wang, *Fellow, IEEE*

***Abstract*—The demand for Text-to-Audio (TTA) generation has grown significantly in domains such as cross-modal creation, intelligent interaction, and accessibility applications. However, existing TTA models suffer from three critical limitations: inadequate modeling of long-range temporal dependencies in extended audio sequences, insufficient restoration of transient acoustic details, and excessively high computational overhead that severely restricts deployment in resource-constrained scenarios. To address these issues, this study proposes a lightweight TTA model named LMafussion, which enables high-quality TTA generation with limited hardware resources. Specifically, this study designs a lightweight Linear Diffusion Module (LDiM) as the diffusion backbone network, replacing traditional structures such as U-Net and UViT. This module integrates the proposed Global Linear Diffusion Transformer (GLDiT) and Locally Optimized Mamba (LOMamba). The former reduces the quadratic complexity of traditional self-attention to a linear scale via linear attention, effectively lowering computational overhead. The latter dynamically adjusts the decay matrix based on a selective state-space model to strengthen the correlation of transient acoustic details. Through their synergistic effect, they not only break the receptive field limitation of traditional structures but also effectively capture the feature correlations of long-sequence audio. LMafussion achieves excellent performance on the public AudioCaps, MusicCaps, and Clotho datasets while significantly reducing computational overhead, demonstrating its high efficiency in resource-constrained scenarios. Some generated examples are available at https://github.com/AOKiHiNa21/LMafussion.**

***Index Terms*—Text-to-audio, Linear Attention, Selective State-Space Model, Diffusion Transformer**

I. INTRODUCTION

TEXT-to-Audio (TTA) generation aims to synthesize diverse audio signals with semantic matching and acoustic authenticity based on text prompts [1]. It serves as a key technology bridging natural language semantics and auditory perception in cross-modal intelligence. With the in-depth integration of AI technology into fields such as multimedia creation, intelligent interaction, and accessibility services, the application scenarios of TTA have gradually expanded from early simple sound effect synthesis to key domains, including cross-modal content enhancement and real-time sound effect generation in virtual environments. However, the training and inference processes of current TTA models generally consume substantial computing resources, restricting their deployment in resource-constrained scenarios. For example, mainstream models, such as AudioGen [2], DiffSound [3], and AudioLDM [4] have high computational costs due to their complex architectures and large parameter scales. Therefore, the coordinated optimization of TTA models' generation quality and resource adaptability is crucial for enhancing user experience and facilitating the practical deployment of these models.

Early TTA methods relied on handcrafted features and struggled to model complex acoustic textures [5]. With the emergence of generative AI technology, WaveNet [6] and SampleRNN [7] adopted end-to-end sequence networks to directly generate raw waveforms or mel-spectrograms. Nevertheless, due to issues such as autoregressive delay caused by sequential generation and dependence on additional supervision, these methods can barely meet the requirements of TTA for generation accuracy. To address these problems, methods including DiffWave [8], and AudioLM [9] have introduced diffusion models. Through adaptation approaches such as noise prediction, latent space diffusion, and combination with U-Net-based latent diffusion, they avoid autoregressive delay, reduce data dependence, support diverse sound effect generation, and improve TTA generation efficiency and quality.

Existing studies have enabled TTA to achieve breakthroughs in generation quality, but they are often still limited in three significant challenges: (1) Difficulties in capturing global correlations for long-sequence audio. The core output of TTA tasks is often tens of seconds of audio, which requires covering extensive feature correlations to ensure temporal consistency [9]. Inadequate capture of long-range information is prone to causing temporal discontinuity and context disconnection [3], [10]. (2) An inherent tension between building global correlations in long-sequence audio modeling and adapting to limited computational resources. Traditional approaches typically depend on complex modeling mechanisms to capture long-range dependencies, which results in substantial computational and memory overhead that is often beyond the capabilities of standard hardware. (3) Bottlenecks in the accurate restoration of transient acoustic details. Auditory authenticity depends on transient features characterized by subtle variations between adjacent frames. When global scene

Biao Yang, Ruiyang Shen, Yang Chen, and Rongrong Ni are with the Wang Zheng Institute of Microelectronics, Changzhou University, Changzhou, 213000 China. Email: yb6864171@cczu.edu.cn, s24060854025@smail.cczu.edu.cn, chenyangcczu@cczu.edu.cn, nirongrong@cczu.edu.cn

Wenwu Wang is with the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK. Email: w.wang@surrey.ac.uk

Corresponding author: Yang Chen (chenyangcczu@cczu.edu.cn)

This work is supported by the General Project of National Natural Science Foundation of China (62576052); Project of Jiangsu Provincial Department of Science and Technology (BK20250969); Changzhou City Applied Basic Project (CJ20240039); Changzhou Leading Innovative Talent Introduction and Cultivation Project (CQ20250044).

modeling fails to accurately capture these fine-grained details, important features may be lost and high-frequency textures become blurred, ultimately degrading task performance [10], [11]. Consequently, a major challenge in deploying TTA under resource-constrained conditions is to develop a technical solution that simultaneously supports effective long-sequence modeling, maintains a lightweight computational architecture, and preserves local detail reconstruction, while still retaining the high-fidelity benefits of the models.

To address the aforementioned issues, this study proposes a lightweight TTA model called LMafussion, where a lightweight Linear Diffusion Module (LDiM) is designed as the diffusion backbone network of LMafussion, integrating the Global Linear Diffusion Transformer (GLDiT) and Locally Optimized Mamba (LOMamba). Specifically, our main contributions in this study are summarized as follows:

1) To address the challenge of modeling global correlations in long audio sequences, we introduce LDiM as a replacement for the conventional diffusion backbone, capitalizing on its strong global modeling capability. LDiM expands beyond receptive field limitations, effectively captures cross-segment acoustic dependencies, and alleviates temporal discontinuity.
2) We propose the GLDiT module, which replaces conventional self-attention with linear attention and achieves linear computational complexity through kernel-based approximation and low-rank enhancement. By substantially reducing the cost of global correlation modeling, GLDiT effectively resolves the tension between capturing global dependencies and operating under limited computational resources.
3) We introduce the LOMamba module, which utilizes a selective state space to dynamically adjust the decay matrix, thereby strengthening the local correlations of transient tokens and effectively mitigating the detail blurring issue.

LMafussion achieves excellent performance not only on the public AudioCaps dataset [12] but also demonstrates strong generalization on the MusicCaps [13] and Clotho [14] datasets, significantly outperforming those of state-of-the-art (SOTA) models such as AudioLDM-L [4] and AudioGen [2]. In addition, the computational overhead of LMafussion is significantly reduced compared with baseline models, offering substantial advantages in resource-constrained scenarios.

The remainder of this paper is organized as follows: Section II provides an overview of related work; Section III introduces preliminaries; Section IV elaborates on LMafussion in detail; and Sections V and VI present the evaluation results and conclusions, respectively.

II. RELATED WORKS

A. Text-to-Audio Generation

Since WaveNet [6] and SampleRNN [7] introduced autoregressive and hierarchical sequence-to-sequence generation mechanisms, end-to-end text-to-audio synthesis has rapidly replaced traditional concatenative and statistical parametric systems. This transformation simplifies the architecture, eliminates concatenation artifacts, and significantly improves generation efficiency and long-term temporal stability. Subsequent models like AudioLM [9] directly model discrete audio tokens through stacked language modeling paradigms, achieving superior long-form coherence and acoustic fidelity across speech, music, and environmental sounds. MusicLM [13] and MusicGen [15] further leverage large-scale self-attention and diffusion/decoder-only architectures for parallel or iterative decoding, enabling high-fidelity modeling of complex text-audio correlations, cross-domain generation (speech, music, sound effects), and rich conditioning capabilities, thereby dramatically expanding the application scope of text-to-audio systems.

With the emergence of diffusion models, WaveGrad [16] performs progressive denoising directly on acoustic features or latent spaces, avoiding the inherent delay of autoregressive generation. AudioLDM [10] combines latent diffusion with contrastive language audio pretraining (CLAP)-based alignment technologies to reduce deviations between generated audio and input text. IMPACT [17] reduces resource consumption and optimizes inference efficiency through parameter compression and mask decoding. Recent works such as Audiobox [18] have also explored unified audio generation with natural language prompts. Tango2 [19] further advances diffusion-based text-to-audio generation through direct preference optimization.

Although these approaches continue to improve sound quality, their underlying architectures still depend on U-Net structures or conventional self-attention mechanisms [20]. In long-audio scenarios, local receptive field limitations and quadratic complexity coexist, and global modeling tends to smooth transient details [10], [11]. Existing improvements only achieve trade-offs by reducing network width [21], expert selection [22], compressing sampling steps [23], or optimizing for specific data constraints [24], without resolving the tripartite trade-off between long-range dependence, local fidelity, and computational efficiency at the architectural level. By designing a lightweight diffusion backbone, this study balances long-range temporal correlation capture and computational resource minimization via an optimized attention mechanism. Meanwhile, it enhances transient detail restoration through targeted local feature enhancement, overcoming this tripartite trade-off limitation architecturally.

B. Diffusion Model

As the current mainstream generative framework, diffusion models characterize real data distribution via the probabilistic mechanism of forward noising and reverse denoising [25]. Efficiency enhancement research focuses on three directions: sampling step compression, network backbone lightweighting, and attention complexity reduction.

For sampling step compression: DDIM [26] eliminates sampling randomness via a deterministic reverse process to reduce iterations; DPM-Solver [23] reduces steps to 10–50 using high-order Ordinary Differential Equation (ODE) discretization; RIN [27] boosts generation accuracy with few steps through optimal noise scheduling. For audio scenarios, DiffRhythm [28] optimizes parameters like noise covariance to fit long-sequence characteristics. However, these models still

use fundamental backbones (e.g., U-Net), failing to balance audio modeling performance and efficient resource use.

In network lightweighting: SnapFusion [29] adopts a dynamic depth strategy for adaptive exit from shallow modules; Δ-DiT [30] achieves hierarchical computation exit via phased fusion to reduce computational load. Yet these optimizations are only validated for image tasks and not adapted to audio generation needs. AudioLDM-S [10] (lightweight audio models) reduce U-Net channels or remove attention heads—while this improves efficiency, it leads to noticeable loss in high-frequency details.

Attention complexity reduction mainly relies on two approaches: sparse routing and linear approximation. DiffMoE [22] reduces FLOPs via Top-k hard routing, but its complexity remains quadratic; RWKV [31] lowers complexity by transmitting temporal states through recursive gating, yet suffers from recursive error accumulation. In contrast, state-space models like S4 [32] and Mamba [4] efficiently capture long-range dependencies due to linear recurrence. Among them, Mamba's selective state update mechanism is proven effective in biomedical image segmentation [33], [34] and medical image segmentation [35], [36].

The above studies have not simultaneously achieved linear complexity and local fidelity in TTA tasks. This study uses kernel approximation-based low-rank linear attention and selective state space in the same module for collaborative denoising, enabling joint optimization of both.

III. PRELIMINARIES

A. Variational Autoencoder (VAE)

The Variational Autoencoder (VAE) [37] is a core model for constructing an efficient latent space in generative tasks, whose goal is to learn low-dimensional continuous latent representations of data and achieve high-quality reconstruction. The VAE consists of an encoder and decoder, both of which are built using stacked convolutional modules and ResNet blocks. In this task, the input to the encoder is a mel-spectrogram $\boldsymbol{M}$ obtained by the Short-Time Fourier Transform (STFT), with dimensions $\boldsymbol{M} \in \mathbb{R}^{F\times N}$, where F represents the number of frequency bins and N is the number of time frames. The encoder maps $\boldsymbol{M}$ to a low-dimensional latent space through layer-wise downsampling, outputting a mean $\boldsymbol{z}_\mu$ and variance $\boldsymbol{z}_\sigma$, both of which have dimensions $\boldsymbol{z}_\mu, \boldsymbol{z}_\sigma \in \mathbb{R}^{C\times \frac{N}{r}\times \frac{F}{r}}$. Here, C is the number of channels in the latent space, and r is the downsampling compression rate. Then, the initial clear latent feature $\boldsymbol{z}_0$ is generated based on the reparameterization trick, as follows:

$$\boldsymbol{z}_0 = \boldsymbol{z}_\mu + \boldsymbol{z}_\sigma \cdot \boldsymbol{\zeta} \tag{1}$$

where $\boldsymbol{\zeta}$ follows a standard Gaussian distribution $\boldsymbol{\zeta} \sim \mathcal{N}(0, \boldsymbol{I})$. Taking $\hat{\boldsymbol{z}}_0$ as input, the decoder reconstructs it into $\hat{\boldsymbol{M}}$ with the same dimension as the original mel-spectrogram through layer-wise upsampling, that is, $\hat{\boldsymbol{M}} \in \mathbb{R}^{F\times N}$, thus completing the mapping loop of "spectrogram - latent feature - spectrogram".

B. Diffusion Model

The diffusion model [25] is a type of probabilistic model that generates data based on a discrete-time Markov chain. Its core lies in modeling the data distribution through the dual processes of forward noise addition and reverse denoising. The forward process defines a gradual degradation process from the original data to pure noise. At each step, Gaussian noise is injected according to a preset noise intensity parameter β_t, as follows:

$$\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon} \tag{2}$$

where $\alpha_t = 1 - \beta_t$, and $\boldsymbol{\epsilon}$ follows a standard Gaussian distribution. Using the cumulative parameter $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, a closed-form expression can be further derived as follows: $\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, which ensures that $\boldsymbol{z}_T$ approximates a standard Gaussian distribution after T steps. The reverse process parameterizes the posterior probability through a neural network and approximates the conditional distribution with a Gaussian distribution, whose formula is as follows:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{z}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{z}_t, t), \sigma_t^2 \boldsymbol{I}) \tag{3}$$

where $\boldsymbol{\mu}_\theta$ is the mean of the state from the previous time step predicted by the model through learning, and σ_t is a fixed or learned variance parameter. During the training phase, the mean squared error between the noise predicted by the model and the actual noise was minimized by simplifying the objective based on the variational lower bound. The loss function is expressed as:

$$L = \mathbb{E}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)\|^2\right] \tag{4}$$

where $\|\cdot\|^2$ denotes the squared L2 norm, $\boldsymbol{\epsilon}$ is the real standard Gaussian noise injected during the forward noise addition process, and $\boldsymbol{\epsilon}_\theta$ is the noise predicted by the neural network with parameters θ.

The core task of TTA generation is to construct a parameterized model $p_\theta(\boldsymbol{x}|\boldsymbol{c})$ based on the text condition $\boldsymbol{c}$, and generate audio signals that are semantically matched with the text and have realistic acoustic characteristics. Here, θ represents the learnable parameters of the model, and $\boldsymbol{x}$ denotes the probability distribution of the original audio.

IV. PROPOSED METHOD

This section presents our proposed lightweight TTA model, i.e. LMafussion, including the diffusion backbone LDiM, the GLDiT based linear modeling, and the LOMamba based detail enhancement.

A. Overview of LMafussion

The entire workflow of the LMafussion model is illustrated in Figure 1. In the training phase, the original audio $\boldsymbol{x}$ is first converted into a Mel-spectrogram $\boldsymbol{M}$ through STFT [38]. Subsequently, the Mel-spectrogram $\boldsymbol{M}$ is encoded into the initial latent feature $\boldsymbol{z}_0$ using a VAE encoder [37]. Next, in accordance with the forward process of the Denoising Diffusion Probabilistic Model (DDPM), Gaussian noise is gradually added to $\boldsymbol{z}_0$ according to a preset noise scheduling strategy, generating noisy latent features $\boldsymbol{z}_t$ at different time steps. Afterward, with the text embedding extracted by the text encoder (CLAP) [39] and the uniformly distributed time step t sampled from the preset range of time steps as conditional

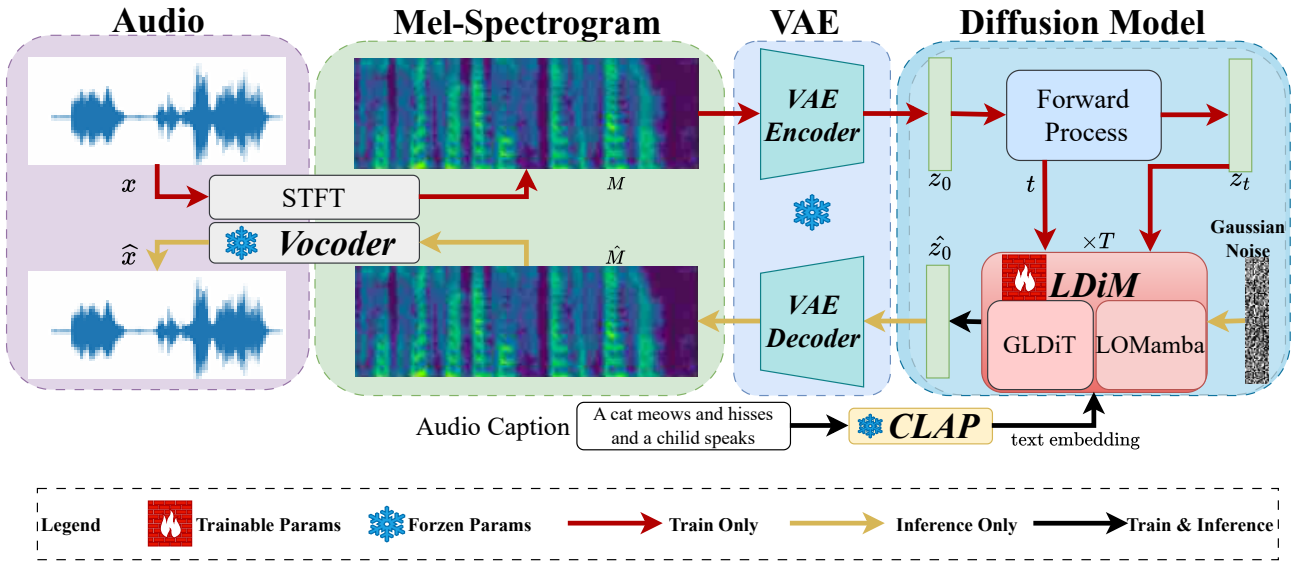

Fig. 1. Overall workflow of LMafussion. This diagram shows the training/inference phases: audio to Mel-spectrogram via STFT, VAE encoding to latent features, forward noising, LDiM denoising with GLDiT&LOMamba under CLAP text embeddings and time conditions, and Vocoder output, balancing efficiency and quality in resource-constrained TTA.

constraints, reverse denoising training is conducted based on the LDiM backbone. The LDiM backbone incorporates two key modules: GLDiT and LOMamba. The former is responsible for efficiently processing long audio sequences to model the long-range temporal correlations, while the latter focuses on enhancing local acoustic details to improve fidelity. These two work together to support the denoising process. At the end of the training phase, model parameters are optimized by calculating the noise prediction loss.

In the inference phase, standard Gaussian noise is used as the initial input. Under the guidance of text embedding, latent features are generated through the inverse diffusion process driven by the LDiM backbone. When the reverse diffusion iterates to $t = 0$, the denoised latent feature $\hat{\boldsymbol{z}}_0$ is obtained; afterward, the VAE decoder reconstructs this latent feature into a Mel-spectrogram $\hat{\boldsymbol{M}}$; finally, the pre-trained Vocoder [40] is used to further convert this spectrogram back into the final time-domain audio signal $\hat{\boldsymbol{x}}$. The following content will specifically introduce the denoising backbone LDiM in LMafussion.

B. LDiM Denoising Backbone

Serving as the primary denoising backbone of the LMafussion model, LDiM addresses the challenge of modeling global correlations in long-duration audio by adhering to the DDPM framework. Through conditional guidance and diffusion-based denoising, it overcomes the inherent limitations of conventional backbone architectures.

*1) **Input Processing**:* The schematic diagram of LDiM is shown in Figure 2. First, patchification processing is performed: non-overlapping patch division (patch size is $p \times p \times C$) is implemented on the 2D latent feature $\boldsymbol{z}_t$; subsequently, each patch is flattened and linearly projected to obtain the initial token sequence $\boldsymbol{z}_{\text{patch}} \in \mathbb{R}^{B \times N \times D}$, where $N = (H/p) \times (W/p)$ is the length of the token sequence, and D is the token feature dimension.

To retain the temporal and frequency position information of audio, LDiM introduces absolute sine-cosine position encoding $\text{pos}_{\text{embed}} \in \mathbb{R}^{N \times D}$, and superimposes this absolute position encoding with the initial token sequence to obtain the token sequence $\boldsymbol{z}'_t$ containing position information. Meanwhile, to achieve the coordinated guidance of temporal and semantic information, the LDiM first obtains the uniformly sampled time step t by sampling from the preset range of time steps during the forward noising process of the diffusion model, and then maps this time step t to the temporal feature $\boldsymbol{t}_{\text{emb}}$ via a Multiple Layer Perceptron (MLP); the text encoding is linearly mapped to the hidden dimension to obtain the text embedding $\boldsymbol{c}_{\text{emb}}$; finally, the two are superimposed to form the global condition $\boldsymbol{c}_{\text{total}} \in \mathbb{R}^{B \times D}$.

*2) **Reverse Denoising**:* The reverse denoising phase is implemented by connecting L prediction blocks in series: the l-th prediction block takes the output $\boldsymbol{z}_t'^{(l-1)}$ of the $(l-1)$-th block and the global condition $\boldsymbol{c}_{\text{total}}$ as inputs, and its output is $\boldsymbol{z}_t'^{(l)}$. After L prediction blocks, the output $\boldsymbol{z}_t'^{(L)}$ is obtained; the linear projection is used to convert $\boldsymbol{z}_t'^{(L)}$ into the noise prediction value $\boldsymbol{\epsilon}_\theta = \text{Proj}(\boldsymbol{z}_t'^{(L)})$, and then $\boldsymbol{\epsilon}_\theta$ is reshaped into the 2D feature $\hat{\boldsymbol{\epsilon}}_\theta$; finally, the noisy latent feature is denoised according to DDPM, with the formula as follows:

$$\hat{\boldsymbol{z}}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{z}'_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_\theta \right) + \sigma_t \cdot \boldsymbol{\epsilon}_t \tag{5}$$

where $\alpha_t = 1 - \beta_t$, σ_t^2 is the noise variance, and $\boldsymbol{\epsilon}_t$ is standard Gaussian noise. $\hat{\boldsymbol{z}}_{t-1}$ is the denoised latent feature corresponding to time step $t-1$ in the reverse process; as the core intermediate output of the reverse denoising iteration, $\hat{\boldsymbol{z}}_{t-1}$ is directly used as the input feature for reverse denoising at the next time step for subsequent denoising calculations.

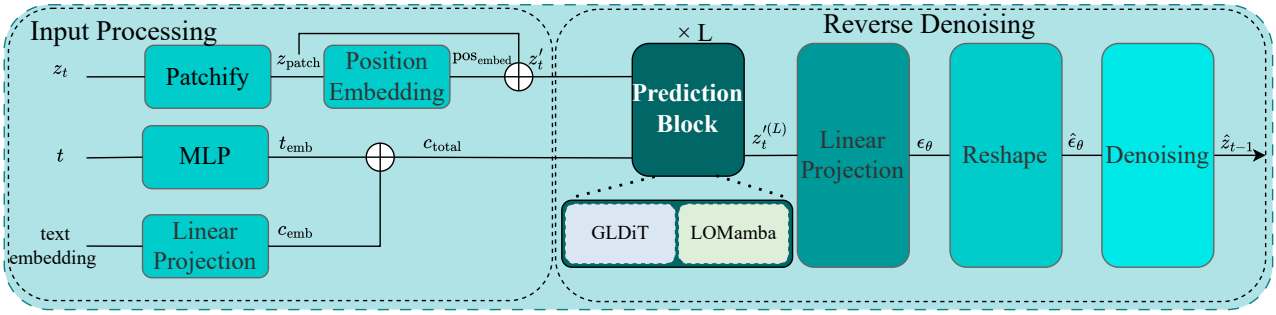

Fig. 2. LDiM backbone schematic. This illustrates the denoising structure replacing U-Net/UViT: patchification, position embedding, AdaLN with text/time conditions, GLDiT and LOMamba blocks for reverse denoising, noise prediction, and feature update, enabling global modeling and detail enhancement for long-audio TTA.

Through this iterative denoising operation, the model can gradually remove noise, enabling the latent representation to approximate the distribution characteristics of real audio data and laying a foundation for the final audio generation from the latent space.

C. GLDiT Sub-module

As the core of global modeling in the LDiM backbone network, GLDiT replaces traditional self-attention by linear attention and combines a positive definite kernel function approximation with a low-rank supplementary strategy to reduce computational complexity to a linear level. This ensures the accuracy of long-range temporal correlation modeling while reducing resource consumption, thereby capturing global information for long-temporal audio generation.

*1) **Adaptive Layer Normalization**:* The schematic diagram of GLDiT is shown in Figure 3. To ensure that the global conditional signal $\boldsymbol{c}_{\text{total}}$ can effectively guide the feature distribution, GLDiT first performs adaptive layer normalization on the input token sequence $\boldsymbol{z}'_t$: it extracts the scaling coefficient $\boldsymbol{\gamma} \in \mathbb{R}^{B\times D}$, offset coefficient $\boldsymbol{\beta} \in \mathbb{R}^{B\times D}$, and gating coefficient $\boldsymbol{\alpha} \in \mathbb{R}^{B\times D}$ from $\boldsymbol{c}_{\text{total}}$ via an MLP, and then performs normalization on the input features based on $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, as follows:

$$\boldsymbol{z}'_{\text{norm1}} = \boldsymbol{\gamma} \odot \frac{\boldsymbol{z}'_t - \mu(\boldsymbol{z}'_t)}{\sqrt{\sigma^2(\boldsymbol{z}'_t) + \delta}} + \boldsymbol{\beta} \tag{6}$$

where $\mu(\cdot)$ and $\sigma^2(\cdot)$ respectively represent the mean and variance of the token sequence in the feature dimension, δ is a small value to avoid division by zero, and $\boldsymbol{z}'_{\text{norm1}} \in \mathbb{R}^{B\times N\times D}$ denotes the normalized feature. This operation enables the feature distribution to dynamically adapt to text semantics and diffusion time steps, providing more robust input for subsequent attention calculation.

*2) **Rotary Position Embedding**:* To enhance the relative positional correlation between tokens, the GLDiT introduces Rotary Position Embedding (RoPE) [41] to perform rotation transformation on $\boldsymbol{z}'_{\text{norm1}}$, yielding $\boldsymbol{z}'_{\text{rot}}$. Through dimension-sensitive rotation angles, RoPE enables differentiated modeling of position sensitivity across different feature dimensions.

*3) **Linear-Augmented Rank-Completed Attention**:* As a core component of GLDiT, we design Linear-Augmented Rank-Completed Attention (LARCA) by replacing traditional self attention with linear attention. In traditional self-attention, token correlations are calculated via the outer product of Query ($\boldsymbol{Q}$) and Key ($\boldsymbol{K}$): $\text{Attn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D}}\right)\boldsymbol{V}$, incurring $O(N^2D)$ time and memory complexity that becomes prohibitive for long audio sequences. To overcome this limitation, LARCA employs linear attention with the non-negative kernel $\phi(\boldsymbol{x}) = \text{elu}(\boldsymbol{x}) + 1$, achieving linear complexity in sequence length. However, as widely observed, the kernel projection on $\boldsymbol{K}$ induces severe rank collapse in the $\boldsymbol{K}\boldsymbol{V}$ cache ($\phi(\boldsymbol{K})^T\boldsymbol{V}$), severely restricting the model's ability to capture rich long-range dependencies in extended audio sequences.

To mitigate this rank collapse while preserving efficiency, LARCA introduces a low-rank augmentation mechanism inspired by the core insight of RALA [42], which first revealed and tackled the rank deficiency problem in kernelized linear attention. Unlike RALA, which restores rank through additional recurrent or adapter-based low-rank modules, LARCA proposes a lightweight residual supplementation strategy that directly injects a low-rank corrective term derived from the original (pre-kernel) $\boldsymbol{K}$ and $\boldsymbol{V}$, formulated as:

$$\begin{gathered}\boldsymbol{Q} = \boldsymbol{W}_Q\boldsymbol{z}'_{\text{rot}},\ \boldsymbol{K} = \boldsymbol{W}_K\boldsymbol{z}'_{\text{rot}},\ \boldsymbol{V} = \boldsymbol{W}_V\boldsymbol{z}'_{\text{rot}} \\ \text{LinAttn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\phi(\boldsymbol{Q})(\phi(\boldsymbol{K})^T\boldsymbol{V})}{\phi(\boldsymbol{Q})\phi(\boldsymbol{K})^T\mathbf{1}} \\ \boldsymbol{z}'_{\text{LARCA}} = \text{LinAttn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) + \boldsymbol{K}\boldsymbol{W}_r(\boldsymbol{V}^T\boldsymbol{W}_r^T)\end{gathered} \tag{7}$$

where $\mathbf{1}$ denotes the all-ones vector and $\boldsymbol{W}_r \in \mathbb{R}^{D\times r}$ ($r \ll D$) is a trainable low-rank projection matrix. This residual low-rank term constructs an independent, high-capacity correlation subspace without compromising kernel non-negativity or reintroducing quadratic costs, maintaining an overall attention complexity of $O(ND^2)$. By combining the efficiency of linear attention with effective rank recovery, LARCA significantly enhances the modeling of complex, long-temporal dependencies critical for high-fidelity text-to-audio generation.

*4) **Feature Enhancement**:* Finally, the gated residual connection and MLP branch work in synergy. First, the output $\boldsymbol{z}'_{\text{LARCA}}$ of LARCA performs a gating operation using the

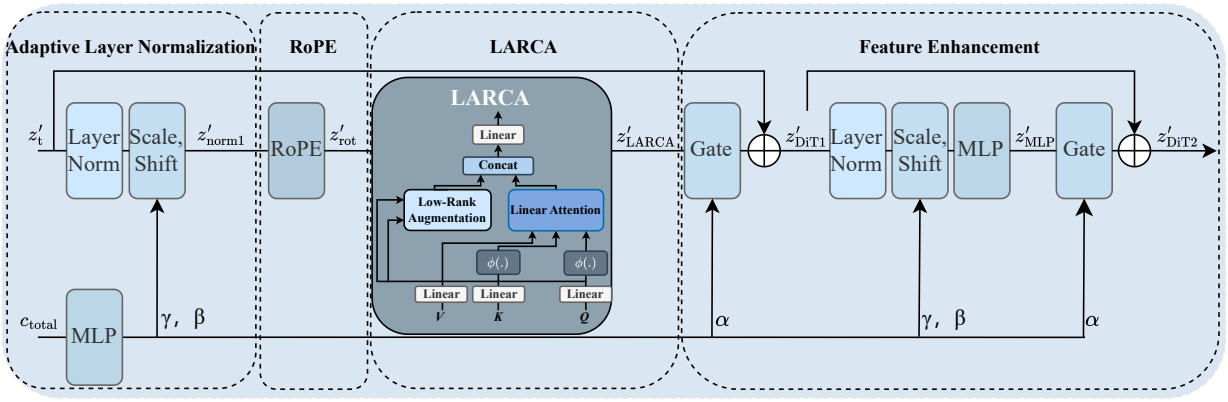

Fig. 3. GLDiT sub-module schematic. This depicts the Global Linear Diffusion Transformer: AdaLN with conditions, RoPE for positions, LARCA (linear attention with low-rank augmentation) for efficient long-range correlations, and gated MLP, reducing complexity while capturing audio dependencies.

gating coefficient $\boldsymbol{\alpha}$ generated from the global condition, and then it is superimposed with the original input $\boldsymbol{z}'_t$ as a residual to obtain $\boldsymbol{z}'_{\text{DiT1}}$; this process can effectively stabilize the gradient in model training. Subsequently, after $\boldsymbol{z}'_{\text{DiT1}}$ is modulated by adaptive layer normalization, it is fed into the MLP with an expansion-compression structure to obtain the local non-linear representation $\boldsymbol{z}'_{\text{mlp}}$. Eventually, $\boldsymbol{z}'_{\text{mlp}}$ is weighted by $\boldsymbol{\alpha}$ and superimposed with $\boldsymbol{z}'_{\text{DiT1}}$ as a residual, outputting the final result $\boldsymbol{z}'_{\text{DiT2}}$ of GLDiT. This operation further enhances the accuracy of acoustic feature expression at the token level, laying a foundation for subsequent local detail optimization.

D. LOMamba Sub-module

Although the GLDiT effectively addresses the core issue of long-sequence processing efficiency, the global attention mechanism tends to exert a smoothing effect on local features, leading to the loss of transient details in audio signals. As the core of local enhancement in the LDiM backbone network, the LOMamba is built based on a selective state-space model. By dynamically adjusting state updates, it achieves accurate capture of local transient correlations between adjacent tokens. It forms a collaborative optimization mechanism of "global long-range modeling - local detail enhancement" with the GLDiT, jointly ensuring long-temporal coherence and the fidelity of local acoustic details.

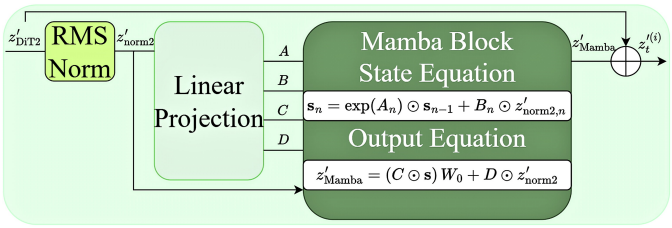

Fig. 4. LOMamba sub-module schematic. This outlines the Locally Optimized Mamba: selective state-space modeling with dynamic decay matrix adjustment for local token correlations, linear projections, and gated residuals, enhancing transient acoustic details in TTA with low overhead.

*1) **Selective State Space Update**:* The schematic diagram of LOMamba is shown in Figure 4. First, Root Mean Square (RMS) normalization is performed on the output $\boldsymbol{z}'_{\text{DiT2}}$ of GLDiT to obtain $\boldsymbol{z}'_{\text{norm2}}$. Next, to realize local correlation modeling, LOMamba converts $\boldsymbol{z}'_{\text{norm2}}$ into four state parameters: $\boldsymbol{A} \in \mathbb{R}^{B\times N\times d}$ (input-dependent decay matrix dynamically generated from fused text-audio features $\boldsymbol{z}'_{\text{norm2}}$), $\boldsymbol{B} \in \mathbb{R}^{B\times N\times d}$ (input projection matrix), $\boldsymbol{C} \in \mathbb{R}^{B\times N\times d}$ (output projection matrix), and $\boldsymbol{D} \in \mathbb{R}^{B\times N\times D}$ (residual bias).

Selective state-space update is the core mechanism of LOMamba. This mechanism achieves progressive state forgetting based on the decay matrix $\boldsymbol{A}$, enabling adjacent tokens to have a greater influence weight on the current state. In contrast, the influence of distant tokens gradually decays, effectively enhancing local feature modeling. The specific formula is as follows:

$$\boldsymbol{s}_n = \exp(\boldsymbol{A}_n) \odot \boldsymbol{s}_{n-1} + \boldsymbol{B}_n \odot \boldsymbol{z}'_{\text{norm2},n} \tag{8}$$

where $\boldsymbol{s}_n \in \mathbb{R}^{B\times d}$ and $\boldsymbol{s}_{n-1} \in \mathbb{R}^{B\times d}$ represent the state vectors of the n-th and $n-1$-th tokens, respectively; $\exp(\boldsymbol{A}_n)$ ensures that the decay coefficient is non-negative to achieve smooth forgetting; $\boldsymbol{B}_n \odot \boldsymbol{z}'_{\text{norm2},n}$ weights and integrates the current token feature into the state vector, enhancing the expression of the current local feature.

*2) **Feature Output**:* To convert local transient correlations into reusable feature representations, LOMamba introduces a linear transformation and feature fusion mechanism in the feature output stage, mapping the local correlation information updated by the state space into feature representations that can be directly used for subsequent processing. The formula is as follows:

$$\boldsymbol{z}'_{\text{Mamba}} = (\boldsymbol{C} \odot \boldsymbol{s})\boldsymbol{W}_o + \boldsymbol{D} \odot \boldsymbol{z}'_{\text{norm2}} \tag{9}$$

where $\boldsymbol{s} = \left[\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N\right] \in \mathbb{R}^{B\times N\times d}$ is the state vector sequence of all tokens, and $\boldsymbol{W}_o \in \mathbb{R}^{d\times D}$ is an output projection matrix, which is used to map the state vector back to the original feature dimensions.

Finally, the locally enhanced features of LOMamba and the global features of GLDiT are superimposed through residual connection. This operation not only retains long-temporal

consistency but also enhances local transient details. The formula is as follows:

$$z_t'^{(i)} = z'_{\text{Mamba}} + z'_{\text{DiT2}} \tag{10}$$

The final output $z_t'^{(i)} \in \mathbb{R}^{B\times N\times D}$ is the denoised feature predicted by the i-th prediction block, which provides high-quality input for subsequent iteration or final noise prediction.

V. EXPERIMENTS AND ANALYSIS

A. Experimental Setup

*1) **Datasets**:* Experiments were performed on the AudioCaps dataset [12], including approximately 49,000 10-second audio clips. Annotated texts cover sound sources, environments, and temporal relationships, and the dataset is divided into training, validation, and testing sets with a 9:0.8:0.2 ratio. All samples were uniformly resampled to 16 kHz mono. Segments shorter than 10 seconds were zero-padded to the standard duration, while those longer than 10 seconds were truncated to 10 seconds.

Mel-spectrogram features were extracted via STFT with specific parameters. These parameters include 2048 window length, 512 step size, 128 Mel channels, and 20–8000 Hz frequency range. The extracted features were subsequently normalized.

To show the generalization capability and generation performance of the proposed model, we also conduct evaluations on the MusicCaps [13] and Clotho [14] datasets. Specifically, MusicCaps is a standard dataset suitable for text-driven music generation, with a test set of about 1,000 audio clips, each of 10 seconds, suitable for assessing music generation quality. Clotho is an audio captioning dataset suitable for generation of general audio and multi-event environmental sounds, with 1,043 environmental audio samples in its test set, each lasting 15–30 seconds, making it ideal for validating long-term and multi-acoustic event modeling capabilities. Notably, the proposed model only employs the test sets of the two aforementioned datasets for zero-shot evaluation, and none of their data is used in the training process.

*2) **Evaluation Metrics**:* Model performance is comprehensively evaluated via objective and subjective metrics as in [4]. Objective metrics include Fréchet Audio Distance (FAD) [50], Fréchet Distance (FD) [51], Inception Score (IS) [52], and Kullback-Leibler (KL) divergence [53]. FAD quantifies auditory feature distribution proximity between generated and real audio. FD reflects latent space feature representation consistency. IS indicates generated audio quality and diversity, with higher values preferable. KL divergence measures distribution differences, with lower values indicating greater similarity [10], [54].

Subjective evaluation is performed by six audio professionals, covering overall audio quality (OVL) and relevance of audio with texts (REL) [43]. OVL assesses audio naturalness, clarity, and listening experience. REL measures audio-text matching degree. A 100-point system is adopted, with higher scores preferable. Results are presented as mean ± standard deviation to ensure reliability [10].

To present results more intuitively in tables, the optimization directions of these metrics are clarified: FD, FAD and KL divergence are better with smaller values (marked as "↓"), while IS, OVL and REL are better with larger values (marked as "↑"). Durations denote the total cumulative duration of audio samples used for model training, measured in hours. It quantifies the scale of training data resources required for model convergence [10].

Real-Time Factor (RTF) is a core metric for evaluating the inference efficiency of audio generation. It is defined as the ratio of the inference time required to generate a given duration of audio to the actual duration of the audio. A smaller RTF value indicates faster inference speed and better real-time performance, making the model more suitable for deployment on resource-constrained devices and for real-time interaction scenarios.

*3) **Experimental Details**:* All experiments are based on the PyTorch framework, using 4 NVIDIA RTX 2080 GPUs. The optimizer adopts AdamW, and the learning rate scheduling strategy is ReduceLROnPlateau.

LDiM stacks 10 prediction blocks. Each layer of GLDiT has a hidden dimension of 1152, with the linear attention kernel function being elu(x) + 1. LOMamba has a state dimension of 256, the epsilon for RMS normalization is set to 1e-6, and the decay matrix required for state update is generated through a 2-layer MLP. Time-step encoding is implemented by combining sine-cosine encoding with a 2-layer MLP. Text conditional encoding uses a pre-trained CLAP model.

*4) **Baseline Methods**:* The proposed method is compared with other SOTA TTA methods detailed below, whose specific configurations are elaborated as follows:

1) The AudioLDM [10] series: the lightweight model AudioLDM-S, the large-parameter model AudioLDM-L, and the upgraded version AudioLDM2 [48];
2) U-Net diffusion-based methods: DiffSound [3], Riffusion [44], TANGO [55], and Auffusion [45];
3) Transformer diffusion-based methods: AudioGen [43], Make-An-Audio [56], Make-An-Audio2 [57], MeanAudio [49], and AudioLCM [46].
4) Text-to-music generation models: MusicGen-Medium [58], MusicLM [13], MeLoDy [59], Mousai [56], AudioLDM2-MSD [48], and Riffusion [44].

These baselines are selected to cover the most representative U-Net and Transformer diffusion architectural paradigms in TTA, as well as the full model spectrum from lightweight to large-scale. As widely recognized standard benchmarks, they ensure high comparability and academic validity of comparisons. Recent works such as TangoFlux [60], GenAU [61], and ETTA [62] adopt fundamentally different training paradigms including flow matching, large-scale synthetic caption scaling, and preference optimization, which are inconsistent with our focus here on architectures. Direct and fair comparison requires alignment with their specific experimental setups which is not trivial, thus they are not included in our comparisons.

To ensure the fairness and transparency of comparative evaluations, all baseline metrics are obtained directly from the original published papers. A unified evaluation protocol is used across all experiments to eliminate discrepancies caused

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT MODELS ON AUDIOCAPS

Model	Params	Durations (h)	FD↓	FAD↓	KL↓	IS↑	OVL↑	REL↑
DiffSound [3]	400M	5420	47.68	7.75	2.52	4.01	45.00±2.6	43.83±2.3
AudioGen [43]	285M	8067	——	3.13	2.09	——	71.68±1.89	66.01±1.79
Riffusion [44]	1.1B	1990	26.28	4.68	1.57	7.21	——	——
Auffusion [45]	1.1B	145	24.45	2.25	1.39	10.14	——	——
AudioLCM [46]	824M	——	26.69	1.74	1.78	8.08	66.71±0.8	67.12±1.4
Make an Audio [47]	543M	3000	——	2.66	1.61	——	——	——
AudioLDM-S [10]	181M	145	29.48	2.43	1.97	6.90	63.41±1.4	64.83±0.9
AudioLDM-L [10]	739M	8886	27.12	2.08	1.86	7.51	64.30±1.6	64.72±1.6
AudioLDM2-full [48]	346M	29510	——	1.78	1.60	7.31	66.48±0.5	65.98±0.9
AudioLDM2-full-L [48]	712M	29510	——	1.86	1.64	7.83	68.27±0.8	68.25±0.8
MeanAudio [49]	**120M**	——	**14.30**	1.77	1.32	10.02	70.03±0.3	70.21±0.6
LMafussion r=64 (ours)	257M	145	24.79	1.54	1.55	9.93	71.24±0.6	70.72±0.7
LMafussion r=128 (ours)	257.3M	145	18.32	1.39	**1.28**	**11.02**	**73.11±0.5**	**72.27±0.6**
LMafussion r=256 (ours)	257.9M	145	18.34	**1.38**	1.30	10.86	72.90±0.5	72.11±0.6

by differences in experimental setup, hardware, or feature-extraction configurations.

B. Quantitative Analysis

As reported in Table I, the LMafussion model has a compact parameter size of only 257.3 million. Experiments were conducted on the AudioCaps dataset. For the metrics of FD (18.32) and FAD (1.39), GLDiT constructs global dependencies for long audio sequences through the DiT architecture to ensure temporal consistency with real audio, while LOMamba optimizes the decay matrix to better capture local acoustic details and mitigate high-frequency blurring. The synergy of these two modules enables LMafussion to outperform AudioGen, which relies on an autoregressive Transformer architecture. This design suffers from inherent error accumulation during long-sequence generation and lacks efficient parallel modeling, resulting in poor temporal coherence and limited ability to capture fine-grained acoustic details. Although MeanAudio exhibits a distinct advantage in the FD metric, it is inferior to LMafussion in almost all other key evaluation indicators.

For IS (11.02) and KL divergence (1.28), GLDiT's low-rank strategy compensates for linear attention loss, balancing quality and diversity. LDiM aligns the distribution of generated audio with that of real audio via text-guided diffusion denoising, achieving better performance than Auffusion and MeanAudio. Auffusion employs a U-Net–based diffusion framework but lacks efficient global modeling for long audio sequences. MeanAudio adopts a mean-flow-based generation architecture, which is incapable of capturing both long-range temporal dependencies and transient acoustic details simultaneously due to its inherent architectural limitations. As a result, both methods underperform compared to LMafussion in terms of balancing quality and diversity and achieving accurate distribution alignment. In subjective evaluation, LOMamba enhances local transient features to improve OVL. GLDiT preserves text-audio semantic correlation via CLAP to ensure stable REL. Eventually, LMafussion achieves an OVL of 73.11 and an REL of 72.27, outperforming Riffusion. This is because Riffusion relies on U-Net diffusion and has no dedicated text-audio semantic matching design. Based on the preceding objective and subjective evaluations, LMafussion's performance scores are predominantly concentrated in the upper range. The model demonstrates strong audio quality, high generation diversity, and close alignment with the target distribution.

Notably, LMafussion is trained exclusively on the 145-hour AudioCaps dataset, matching the training data scale of the lightweight baseline AudioLDM-S. This controlled setup rigorously validates the lightweight superiority of our architecture: it removes data scale as a confounding factor, avoids the overhead of large-scale pre-training, and confirms that performance improvements stem mainly from architectural advances rather than from expanded training data.

C. Sensitivity Analysis of Hyperparameters

To explore the impact of critical hyperparameters on model performance, we perform a sensitivity analysis of the low-rank dimension r of the LARCA module. Three representative values are evaluated: $r = 64$, $r = 128$, and $r = 256$, respectively.

As reported in Table I, increasing the low-rank dimension from 64 to 128 yields substantial performance gains across multiple metrics (e.g., FD and FAD). This phenomenon demonstrates that a proper increase in r effectively mitigates the rank collapse issue in linear attention, thereby strengthening feature representation and global modeling for long-sequence audio. When r is further increased to 256, the overall model performance reaches a plateau, with no significant improvement, but slight degradation observed in some metrics. This suggests that an excessively large low-rank dimension may introduce redundant feature representations, thereby impairing the model's generalization capability and training stability.

In comparison, $r = 128$ strikes a good trade-off between audio generation quality, long-term temporal modeling capacity, and computational complexity, validating the rationality of the hyperparameter configuration adopted in this work.

D. Experimental Results on MusicCaps

This section compares the proposed model with several state-of-the-art text-to-music generation models on the

MusicCaps dataset, including MusicGen-Medium [58], MusicLM [13], MeLoDy [59], Mousai [56], AudioLDM2-MSD [48], and Riffusion [44]. The results of all the compared models are adopted from their original papers or official open-source implementations.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT AUDIO GENERATION MODELS ON THE MUSICCAPS DATASET

Model	FAD↓	KL↓	OVL↑	REL↑
Riffusion [44]	14.80	2.06	-	-
Mousai [56]	7.50	1.59	-	-
MeLoDy [59]	5.41	-	-	-
MusicLM [13]	4.00	-	-	-
MusicGen-Medium [58]	4.89	1.35	67.40	67.60
AudioLDM2-MSD [48]	4.47	1.32	68.20	66.00
LMafussion (ours)	**3.25**	**1.31**	**70.12**	**69.23**

As reported in Table II, the proposed LMafussion method achieves superior performance on all key metrics. Concretely, LMafussion obtains a FAD score of 3.25, which is approximately 27.3% lower than that of AudioLDM2-MSD (4.47) and 33.5% lower than that of MusicGen-Medium (4.89). For the KL divergence metric, LMafussion achieves a value of 1.31, outperforming AudioLDM2-MSD (1.32) and MusicGen-Medium (1.35), demonstrating stronger text-audio feature alignment capability. In terms of the OVL and REL scores, which are highly correlated with text-semantic matching, LMafussion reaches 70.12 and 69.23, respectively, achieving the best results among all the compared models.

These results indicate that learning the audio generation models from a general perspective can also deliver excellent zero-shot performance in the specific music domain, demonstrating the superiority and cross-domain generalization capability of the proposed general framework.

E. Long-Audio Generation Performance on Clotho

To further validate the effectiveness of the proposed model in long-sequence audio generation, comparative experiments are conducted on the Clotho dataset, designed for long-duration, multi-event environmental sounds. Several mainstream models, including TANGO [55], AudioLDM-L [10], and Make-An-Audio2 [57], are selected as baselines to evaluate long-sequence modeling, temporal consistency, and detail fidelity.

TABLE III
PERFORMANCE COMPARISON ON THE CLOTHO DATASET

Model	FD↓	IS↑	KL↓	FAD↓
TANGO [55]	32.10	6.77	2.59	3.61
AudioLDM-L [10]	28.15	6.55	2.60	4.93
Make-An-Audio2 [57]	**19.97**	8.50	2.49	**2.13**
LMafussion (ours)	21.82	**9.48**	**2.46**	2.57

As reported in Table III, LMafussion exhibits competitive overall performance, comparable to that of state-of-the-art methods, in long-audio generation tasks. Specifically, the proposed model achieves a KL divergence of 2.46, surpassing all comparative methods, thereby verifying its stable text-audio semantic alignment in long-duration and multi-acoustic-event scenarios. For IS, LMafussion attains 9.48, outperforming Make-An-Audio2 (8.50) and AudioLDM-L (6.55), demonstrating its ability in preserving richer acoustic details and higher sample diversity in long-audio generation. Regarding the distribution-matching metrics, FD and FAD, LMafussion achieves 21.82 and 2.57, respectively, which are slightly inferior to the top-performing Make-An-Audio2 but comparable to this method, and significantly outperform TANGO and AudioLDM-L. This demonstrates the outstanding distribution fitting ability and temporal coherence of the proposed model in long-sequence audio modeling.

The above results fully demonstrate that the proposed GLDiT linear long-range modeling mechanism can effectively overcome the limitations of receptive field constraints and temporal discontinuities in traditional models for long-audio processing, enabling high-quality, high-consistency long-audio generation even under resource-constrained conditions.

F. Qualitative Analysis

To intuitively validate LMafussion's advantages in local detail restoration and temporal logic capture, this section compares spectral feature differences between audio generated by LMafussion and the baseline AudioLDM-L.

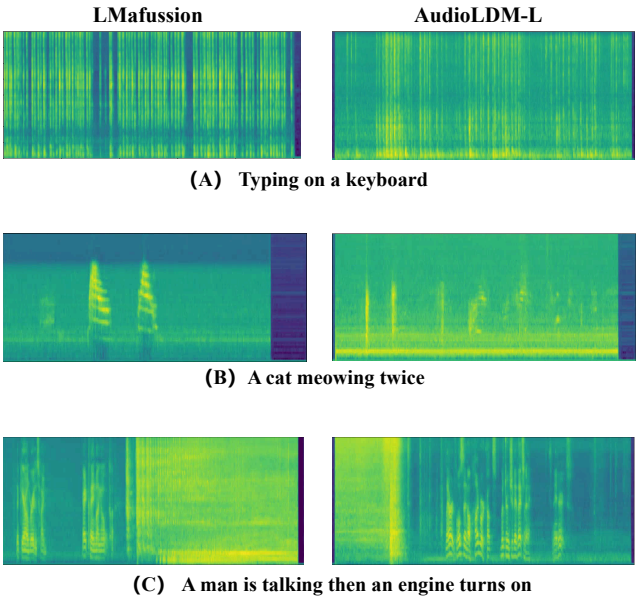

Fig. 5. Visual comparison of the generated audio by LMafussion and AudioLDM-L.

As illustrated in (A) of Figure 5, the "keyboard typing sound" task shows that AudioLDM-L's spectral texture is blurred, with noise-induced energy peak dispersion, resulting in the loss of typing's sharp texture. LMafussion effectively models discrete impact features, clearly separating high-frequency impulses from low-frequency background noise. This capability arises from tokenization that preserves local spectral independence, combined with LOMamba's dynamic decay matrix, which captures transient local details under text-semantic guidance.

In (B) of Figure 5, the "a cat meowing twice" task shows AudioLDM-L's two meow energy peaks are blurred, with low discriminability, failing to capture event discreteness. LMafussion exhibits two independent sharp peaks with distinct time intervals. This is attributed to GLDiT's LARCA mechanism solving linear attention limitations and LOMamba dynamically adjusting state matrices to clarify temporal boundaries, leveraging their collaborative advantage in temporal detail modeling.

In (C) of Figure 5, the "a human voice followed by engine startup" task demonstrates that AudioLDM-L has blurred the boundaries between voice and engine sound, with an ambiguous temporal logic. LMafussion generates a clear first-half voice and well-stratified second-half engine sound, with natural temporal transition. This is due to DiT's global self-attention modeling long-range dependencies and GLDiT's rotary positional encoding enhancing time step discriminability, supported by the model's innovative architecture for temporal sequence coherence.

G. FD Index Convergence Comparison Analysis

To verify LMafussion's collaborative advantages in terms of training efficiency and distribution fitting capability, this section analyzes its FD convergence characteristics. It is compared with DiT-based models, such as CrossDiT [63], PixelArt-DiT [30], Stable-Audio-DiT [64], and Ezaudio [21].

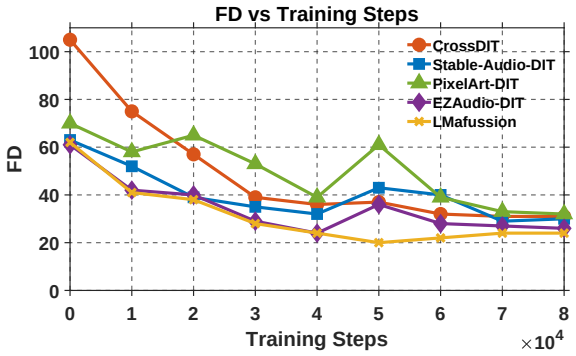

Fig. 6. Comparison of convergence characteristics of different models.

As illustrated in Figure 6, LMafussion's FD dropped sharply from 62 to 41 in early training within 10,000 steps. Its reduction was significantly greater than CrossDiT, PixelArt-DiT, and Stable-Audio-DiT, and only slightly better than Ezaudio-DiT. At 50,000 steps, its FD stabilized at 20–24. Ezaudio-DiT stayed at 26–28, and Stable-Audio-DiT fluctuated at 29–30. This advantage comes from the innovative synergy of GLDiT and LOMamba in LMafussion: GLDiT uses global self-attention with linear complexity to build long-term audio correlations and capture distribution profiles efficiently, while LOMamba dynamically adjusts decay matrices to optimize distribution details in late training, stabilizing FD at low values. Other models lack such integrated local enhancement and efficiency optimization, leading to slower FD convergence speed and higher final stable values than LMafussion.

H. GFlops Comparison Analysis

To quantitatively assess LMafussion's computational efficiency, this section used single-step denoising floating-point operations (GFlops) as an indicator and compared it with TTA generation models of similar or larger parameter scales. As illustrated in Figure 7, LMafussion's single-step denoising GFlops was 176.93, slightly higher than AudioLDM-M-full (415M parameters) due to long-sequence global modeling needs, but much lower than Ezaudio (596M parameters). Stable-Audio-DiT (DiT-based) showed lower long-term efficiency than LMafussion due to a lack of optimization. In terms of distribution of the computational loads across the modules, the GLDiT module equipped with the LARCA mechanism accounts for approximately 64% of the total GFlops, and the LOMamba module accounts for about 33%. It is evident that the computational overhead of the model is mainly concentrated in the GLDiT module, which is responsible for global long-sequence modeling. This advantage originates from the core module innovations: GLDiT converts self-attention's quadratic complexity to linear via the proposed LARCA mechanism, reducing long-sequence token computation overhead, and LOMamba integrates hardware-aware optimizations (parallel scanning, kernel fusion) through selective state space models' linear time processing, cutting local processing costs. Other models rely on high-complexity attention or lack such collaborative optimization, limiting their efficiency in long/high-sampling audio scenarios. LMafussion thus achieves an innovative balance of efficiency, global modeling, and local quality to support real-time TTA generation.

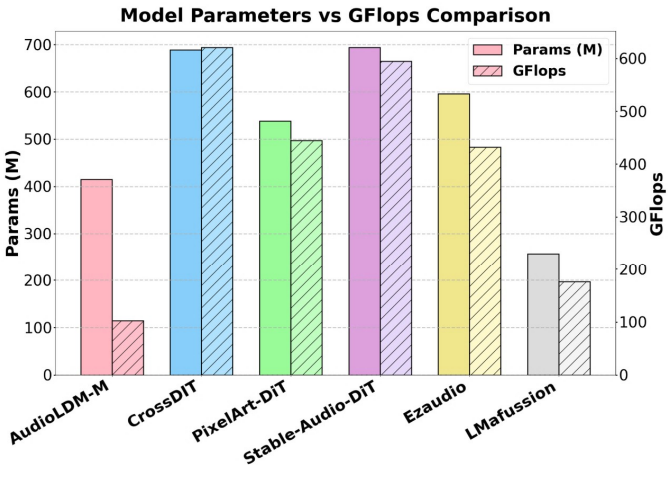

Fig. 7. Comparison of floating-point operations of different models.

I. RTF Analysis

As illustrated in Figure 8, this paper quantitatively compares the RTF of LMafussion with those of the state-of-the-art TTA models. On a single NVIDIA RTX 2080 GPU, the RTF of

TABLE IV
PERFORMANCE COMPARISON OF MODELS WITH DIFFERENT COMPONENT CONFIGURATIONS

Model	FD↓	FAD↓	KL↓	IS↑	OVL↑	REL↑
AudioLDM-S [10]	29.48	2.43	1.97	6.90	63.41±1.4	64.83±0.9
Only DiT	26.24	1.91	1.80	8.12	65.11±1.4	65.38±1.5
With LOMamba	25.73	1.60	1.61	9.65	67.36±0.7	68.23±0.9
With GLDiT	24.9	1.87	1.80	9.67	66.42±1.9	64.36±1.7
With Linear	28.18	1.69	1.85	8.54	66.85±1.2	67.14±1.1
With GRU	32.75	1.81	1.75	8.74	65.92±1.5	66.37±1.3
With RALA	26.79	1.89	1.56	8.67	67.60±1.0	67.95±0.8
LMafussion (ours)	**18.32**	**1.39**	**1.28**	**11.02**	**73.11±0.5**	**72.27±0.6**

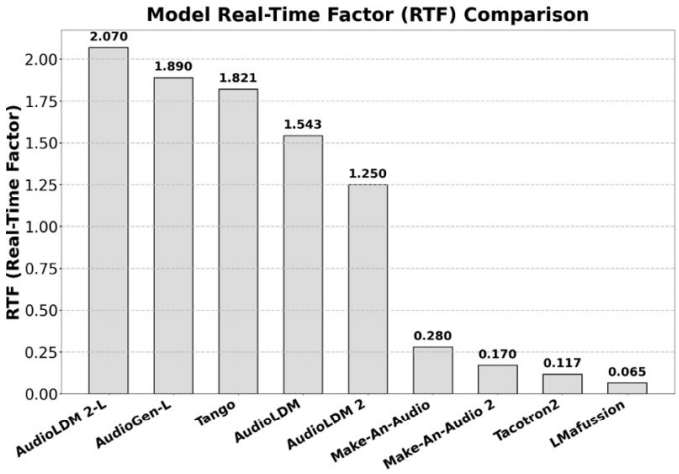

Fig. 8. Comparison of RTF (Real-Time Factor) of different models.

LMafussion is as low as 0.065, which is far superior to those of the baseline models such as AudioLDM2-L (1.543), AudioGen (1.890), and TANGO (1.821). Benefiting from the GLDiT module, which reduces self-attention complexity to linear complexity, and the efficient parallel modeling of local features by LOMamba, LMafussion achieves extremely fast inference speed and can meet the stringent latency requirements of real-time interaction and edge deployment.

J. Ablation Study

This section performs ablation experiments on the AudioCaps dataset to verify the effectiveness of each core component of the LMafussion model. Using AudioLDM-S as the baseline model, we set up multiple comparative variants: "Only DiT" (retaining only the basic DiT architecture), "With GLDiT" (replacing self-attention with the proposed GLDiT module), "With LOMamba" (incorporating the LOMamba local optimization module), "With Linear" (adopting ordinary linear attention without low-rank enhancement), "With GRU" (using GRU instead of LOMamba for local modeling), and "With RALA" (replacing the LARCA module with RALA linear attention).

As shown in Table IV, compared with AudioLDM-S, "Only DiT" achieves significant improvements in FD, FAD, IS, OVL and REL metrics, validating the superiority of the DiT architecture in modeling long-term temporal dependencies. However, constrained by the quadratic complexity of multihead self-attention, it suffers from high memory consumption and tends to produce blurred local transient details. By virtue of the linear complexity of the LARCA module, "With GLDiT" alleviates the rank deficiency problem of conventional linear attention while maintaining favorable temporal consistency. Nevertheless, due to the lack of a targeted local detail enhancement mechanism, its FD, FAD and OVL metrics are inferior to those of the "With LOMamba" variant.

"With LOMamba" achieves excellent performance in both objective metrics and subjective evaluation through selective state space modeling, verifying its capability to recover fine-grained local acoustic details. However, this variant does not optimize the attention complexity, resulting in high memory usage.

To further validate the rationality of our design choices, we conduct additional comparative experiments. "With Linear" underperforms GLDiT across most metrics, especially FD (28.18) and IS (8.54), demonstrating that the low-rank enhancement mechanism in LARCA is crucial for mitigating the rank collapse of linear attention and preserving representation capability. "With GRU" yields a high FD value of 32.75 and generally inferior audio-quality metrics, indicating that GRU-based recurrent modeling struggles to capture transient local correlations in audio, and that the selective state-update mechanism of LOMamba is superior to generic recurrent structures for local modeling. "With RALA" performs worse than GLDiT in FD and FAD metrics, indicating that our LARCA module with lightweight residual low-rank compensation outperforms RALA in modeling audio sequences.

Ultimately, LMafussion achieves improved performance through the collaboration of DiT, GLDiT and LOMamba. GLDiT enables efficient long-sequence modeling with linear complexity, while LOMamba enhances the fidelity of text-aligned local details. Their synergy breaks the triangular trade-off between long-range dependence, local accuracy and computational efficiency, and delivers the triple advantages of efficiency, global modeling and local fidelity.

VI. CONCLUSION

We have introduced LMafussion, a lightweight model designed to address key challenges in text-to-audio generation, including long-range audio dependency modeling, the trade-off between global correlation construction and limited computational resources, and the preservation of transient acoustic details. Built upon the DiT-based lightweight backbone LDiM, the model incorporates two dedicated modules: GLDiT, which reduces the complexity of conventional self-attention to linear

scale through kernel-approximated low-rank linear attention for improved efficiency, and LOMamba, which leverages selective state-space modeling to strengthen transient feature relationships and alleviate detail blurring. Experimental results on the AudioCaps dataset demonstrate that LMafussion outperforms representative baseline methods, confirming that its lightweight design effectively reconciles long-term coherence with fine-grained acoustic fidelity under resource constraints.

Despite these advantages, the current work is limited in the diversity of audio generation scenarios it supports. The model is primarily optimized for environmental sounds and sound effects, with limited adaptability to music-oriented audio and no support for multi-channel generation, which restricts its applicability in more complex auditory contexts. Future research will focus on two main directions. First, we aim to broaden the range of supported audio types by incorporating music-specific acoustic representations and enhancing VAE and vocoder architectures, enabling music generation and multi-channel outputs. Second, we will further improve inference efficiency through advanced sampling strategies and hardware-aware optimization of core modules, reducing latency on edge devices and facilitating real-time interactive applications.

REFERENCES

[1] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," *IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2021.

[2] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *International Conference on Machine Learning*, 2020.

[3] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "DiffSound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2022.

[4] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *ArXiv*, vol. abs/2312.00752, 2023.

[5] J. H. Mcdermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, pp. 926–940, 2011.

[6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Speech Synthesis Workshop*, 2016.

[7] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. M. R. Sotelo, A. C. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *ArXiv*, vol. abs/1612.07837, 2016.

[8] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," *ArXiv*, vol. abs/2009.09761, 2020.

[9] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "AudioLM: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2022.

[10] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. P. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *International Conference on Machine Learning*, 2023.

[11] W. S. Peebles and S. Xie, "Scalable diffusion models with transformers," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2022.

[12] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[13] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "MusicLM: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[14] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[15] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. D'efossez, "Simple and controllable music generation," *ArXiv*, vol. abs/2306.05284, 2023.

[16] T. L. Paine, M. Wang, H. Mei, S. Levine, D. Kumar, Y. Chen, S. Ravuri, V. Sindhwani, O. Vinyals, and N. Parmar, "WaveGrad: Estimating gradients for fast neural audio synthesis," *arXiv preprint arXiv:2009.00713*, 2020.

[17] K.-P. Huang, S.-W. Yang, H. Phan, B.-R. Lu, B. Kim, S. Macha, Q. Tang, S. Ghosh, H. yi Lee, C.-C. Kao, and C. Wang, "IMPACT: Iterative mask-based parallel decoding for text-to-audio generation with diffusion modeling," *ArXiv*, vol. abs/2506.00736, 2025.

[18] 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.

[19] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, "Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 564–572.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems*, 2017.

[21] J. Hai, Y. Xu, H. Zhang, C. Li, H. Wang, M. Elhilali, and D. Yu, "EzAudio: Enhancing text-to-audio generation with efficient diffusion transformer," *ArXiv*, vol. abs/2409.10819, 2024.

[22] M. Shi, Z. Yuan, H. Yang, X. Wang, M. Zheng, X. Tao, W. Zhao, W. Zheng, J. Zhou, J. Lu, P. Wan, D. Zhang, and K. Gai, "DiffMoE: Dynamic token selection for scalable diffusion transformers," *ArXiv*, vol. abs/2503.14487, 2025.

[23] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *ArXiv*, vol. abs/2206.00927, 2022.

[24] X. Zhou, M. Zhang, Y. Zhou, Z. Wu, and H. Li, "Accented text-to-speech synthesis with limited data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1699–1711, 2023.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, vol. abs/2006.11239, 2020.

[26] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *ArXiv*, vol. abs/2010.02502, 2020.

[27] T. Chen, "On the importance of noise scheduling for diffusion models," *ArXiv*, vol. abs/2301.10972, 2023.

[28] Z. Ning, H. Chen, Y. Jiang, C. Hao, G. Ma, S. Wang, J. Yao, and L. Xie, "DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion," 2025.

[29] Y. Li, H. Wang, Q. Jin, J. Hu, P. Chemerys, Y. Fu, Y. Wang, S. Tulyakov, and J. Ren, "SnapFusion: Text-to-image diffusion model on mobile devices within two seconds," *ArXiv*, vol. abs/2306.00980, 2023.

[30] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, 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," *ArXiv*, vol. abs/2310.00426, 2023.

[31] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, M. Grella, G. Kranthikiran, X. Du, X. He, H. Hou, P. Kazienko, J. Kocoń, J. Kong, B. Koptyra, H. Lau, K. S. I. Mantri, F. Mom, A. Saito, X. Tang, B. Wang, J. S. Wind, S. Wozniak, R. Zhang, Z. Zhang, Q. Zhao, P. Zhou, J. Zhu, and R. Zhu, "RWKV: Reinventing rnns for the transformer era," in *Conference on Empirical Methods in Natural Language Processing*, 2023.

[32] E. Nguyen, K. Goel, A. Gu, G. W. Downs, P. Shah, T. Dao, S. A. Baccus, and C. Ré, "S4ND: Modeling images and videos as multidimensional signals using state spaces," *ArXiv*, vol. abs/2210.06583, 2022.

[33] J. Ma, F. Li, and B. Wang, "U-Mamba: Enhancing long-range dependency for biomedical image segmentation," *ArXiv*, vol. abs/2401.04722, 2024.

[34] B. Yang, S. Lu, S. Shen, C. Hu, and R. Ni, "Mmflow: Multimodal pedestrian trajectory prediction based on Mambaformer and one-step mean flow," *Expert Systems with Applications*, vol. 319, p. 132057, 2026.

[35] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "SegMamba: Long-range sequential modeling mamba for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[36] R. Ni, Y. Guo, B. Yang, Y. Liu, H. Wang, and C. Hu, "Mul-vmamba: Multimodal semantic segmentation using selection-fusion-based vision-Mamba," *Knowledge-Based Systems*, vol. 334, p. 115119, 2026.

[37] R. Lopez, P. Boyeau, N. Yosef, M. I. Jordan, and J. Regier, "Auto-encoding variational bayes," 2020.

[38] J. B. Allen and L. R. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, pp. 1558–1564, 1977.

[39] 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," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022.

[40] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *ArXiv*, vol. abs/2010.05646, 2020.

[41] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *ArXiv*, vol. abs/2104.09864, 2021.

[42] S. Zhai, H. Dong, Y. Chen, Q. Liu, X. Li, and J. Wu, "RALA: Recurrent rank-one linear attention for long-sequence modeling," 2024, v1: May 2024.

[43] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. D'efossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually guided audio generation," *ArXiv*, vol. abs/2209.15352, 2022.

[44] S. Forsgren and H. Martiros, "Riffusion: Stable diffusion for real-time music generation," 2022, accessed: 2025-12-08.

[45] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4700–4712, 2024.

[46] H. Liu, R. Huang, Y. Liu, H. Cao, J. Wang, X. Cheng, S. Zheng, and Z. Zhao, "AudioLCM: Efficient and high-quality text-to-audio generation with minimal inference steps," *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

[47] R. Huang, J.-B. 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," *ArXiv*, vol. abs/2301.12661, 2023.

[48] H. Liu, Q. Tian, Y. Yuan, X. Liu, X. Mei, Q. Kong, 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*, vol. 32, pp. 2871–2883, 2023.

[49] X. Li, J. Liu, Y. Liang, Z. Niu, W. Chen, and X. Chen, "MeanAudio: Fast and faithful text-to-audio generation with mean flows," *arXiv preprint arXiv:2508.06098*, 2025.

[50] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet Audio Distance: A reference-free metric for evaluating music enhancement algorithms," in *Interspeech*, 2019.

[51] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," 2024.

[52] C. Donahue, M. Lee, and P. Liang, "The inception score for audio: Using pre-trained classifiers to measure diversity and fidelity in generated audio," 2019, workshop on Machine Learning for Music, ICML 2019.

[53] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "HiFi-Codec: Group-residual vector quantization for high fidelity audio codec," *ArXiv*, vol. abs/2305.02765, 2023.

[54] 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, 2019.

[55] G. Deepanway, M. Navonil, M. Ambuj, and P. Soujanya, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

[56] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, "Mousai: Text-to-music generation with long-context latent diffusion," *arXiv preprint arXiv:2301.11757*, 2023.

[57] Z. Pang, Z. Zhang, Y. Leng *et al.*, "Make-An-Audio 2: Towards real-time high-fidelity audio generation with improved diffusion transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 3974–3988, 2024.

[58] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in neural information processing systems*, vol. 36, pp. 47 704–47 720, 2023.

[59] M. W. Lam, Q. Tian, T. Li, Z. Yin, S. Feng, M. Tu, Y. Ji, R. Xia, M. Ma, X. Song *et al.*, "Efficient neural music generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 17 450–17 463, 2023.

[60] C.-Y. Hung, N. Majumder, Z. Kong, A. Mehrish, A. A. Bagherzadeh, C. Li, R. Valle, B. Catanzaro, and S. Poria, "TangoFlux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," *arXiv preprint arXiv:2412.21037*, 2024.

[61] M. Haji-Ali, W. Menapace, A. Siarohin, G. Balakrishnan, and V. Ordonez, "Taming data and transformers for audio generation," *International Journal of Computer Vision*, vol. 134, no. 3, p. 87, 2026.

[62] S.-G. Lee, Z. Kong, A. Goel, S. Kim, R. Valle, and B. Catanzaro, "Elucidating the design space of text-to-audio models," *arXiv preprint arXiv:2412.19351*, 2024.

[63] 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," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22 669–22 679, 2022.

[64] Z. Evans, J. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Long-form music generation with latent diffusion," in *International Society for Music Information Retrieval Conference*, 2024.