Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

QA-MDT: Quality-aware Masked Diffusion Transformer for Enhanced Music Generation

Chang Li[1], **Ruoyu Wang**[1], **Lijuan Liu**[1], **Jun Du**[1], **Yixuan Sun**[1], **Zilu Guo**[1], **Zhengrong Zhang**[1], **Yuan Jiang**[1], **Jianqing Gao**[2], **Feng Ma**[2]
[1]University of Science & Technology of China, Hefei, China
[2]iFlytek AI Research, Hefei, China
{lc_lca, wangruoyu}@mail.ustc.edu.cn,

Abstract

Text-to-music (TTM) generation, which converts textual descriptions into audio, opens up innovative avenues for multimedia creation. Achieving high quality and diversity in this process demands extensive, high-quality data, which are often scarce in available datasets. Most open-source datasets frequently suffer from issues like low-quality waveforms and low text-audio consistency, hindering the advancement of music generation models. To address these challenges, we propose a novel quality-aware training paradigm for generating high-quality, high-musicality music from large-scale, quality-imbalanced datasets. Additionally, by leveraging unique properties in the latent space of musical signals, we adapt and implement a masked diffusion transformer (MDT) model for the TTM task, showcasing its capacity for quality control and enhanced musicality. Furthermore, we introduce a three-stage caption refinement approach to address low-quality captions' issue. Experiments show state-of-the-art (SOTA) performance on benchmark datasets including MusicCaps and the Song-Describer Dataset with both objective and subjective metrics. Demo audio samples are available at https://qa-mdt.github.io/, code and pretrained checkpoints are open-sourced at https://github.com/ivcylc/OpenMusic/.

1 Introduction

Text-to-music (TTM) generation aims to transform textual descriptions of emotions, style, instruments, rhythm, and other aspects into corresponding music segments, providing new expressive forms and innovative tools for multimedia creation. According to scaling law principles [Peebles and Xie, 2023; Li *et al.*, 2024a], effective generative models require a large volume of training data. However, unlike image generation tasks [Chen *et al.*, 2024a; Rombach *et al.*, 2021], acquiring high-quality music data often presents greater challenges, primarily due to copyright issues and the need for professional hardware to capture high-quality music. These factors make building a high-performance TTM model particularly difficult.

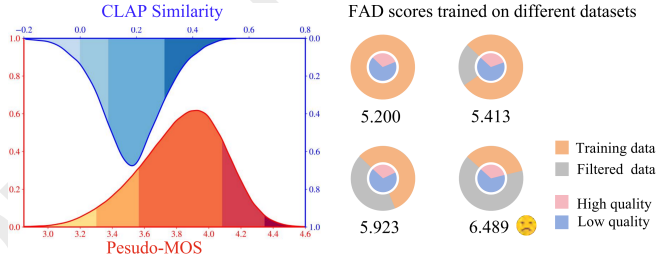

Figure 1: **(a)** The distribution curves of CLAP similarity and pseudo-MOS for large-scale open-source music databases AudioSet [Defferrard *et al.*, 2016] and FMA [Defferrard *et al.*, 2016], where darker areas represent higher text-audio consistency or audio quality. **(b)** Frechet Audio Distance (FAD) [Kilgour *et al.*, 2018] scores on the MusicCaps test set obtained from models trained for 50K steps on AudioSet and FMA, using filter ratios of 0%, 33%, 66%, and 100% of low-quality data. Here, low-quality data is determined by a Pseudo-MOS score [Ragano *et al.*, 2023] of less than 4.0. It can be inferred that performance consistently worsens with larger filter ratios.

In the TTM field, high-quality music signals is scarce. This prevalent issue of low-quality data, highlighted in Figure 1 (a), manifests in two primary challenges. Firstly, most available music signals often suffer from distortion due to noise, low recording quality, or outdated recordings, resulting in diminished generated quality, as measured by pseudo-MOS scores from quality assessment models [Ragano *et al.*, 2023]. Secondly, there is a weak correlation between music signals and captions, characterized by missing, weak, or incorrect captions, leading to low text-audio similarity, which can be indicated by CLAP scores [Wu *et al.*, 2023]. These challenges, especially the **inherent quality in the music signal itself**, significantly hinder the training of high-performance music generation models, resulting in poor rhythm, noise, and inconsistencies with textual control conditions in the generated audio. Additionally, as shown in Figure 1 (b), directly filtering low-quality music, which robustly reduces the dataset size, leads to a consistent decline in model performance. Therefore, finding an effective training strategy for large-scale datasets with low-quality waveforms, mismatches, and missing labels has become an urgent challenge.

In this paper, we introduce a novel quality-aware masked

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

diffusion transformer (QA-MDT) to enhance music generation, aiming to tackle the aforementioned problems while making further improvements through architectural studies. We made efforts on effectively leveraging extensive open-source music databases, which often contain data of varying quality and style, to produce high-quality, diverse and high text-audio consistency music. **For music quality enhancement**, we innovatively inject music quality into the denoising stage with multiple granularities to foster quality awareness during training, while high-quality music can be obtained by setting a quality threshold during inference. **Regarding the modeling architecture**, in preliminary experiments, we found that the Diffusion Transformer (DiT) framework, which has been successful in the image domain [Peebles and Xie, 2023], is not directly effective for modeling music spectrograms. However, injecting a masking strategy significantly enhances the spatial correlation of the music spectrum and further accelerates convergence. Additionally, we utilize large language models (LLMs) and the CLAP model to synchronize music signals with captions, thereby **enhancing text-audio correlation** in extensive music datasets. Our ablation studies on public datasets confirm the effectiveness of our methodology, with the final model surpassing previous works in both objective and subjective measures. In summary, we focus on developing better training strategies and network architectures to enhance the quality and aesthetic of music generation. At the same time, we address the long-standing issue of text-audio consistency in the field of TTM, which can be listed as:

- We propose a quality-aware training paradigm that enables the model to perceive the quality of the dataset during training, thereby achieving superior music generation in terms of both musicality and audio quality.
- We innovatively introduced the Masked Diffusion Transformer to music signals, demonstrating its unique efficacy in modeling music latent space and its capability in perceiving quality control, thereby further improving both the generated quality and musicality.
- We address the issue of low text-audio correlation in large-scale music datasets for TTM, effectively improving text alignment and generative diversity.

2 Related Work

Text to music generation. Text-to-music generation aims to create music clips that correspond to input descriptive or summary text. Previous efforts have utilized either language models (LMs) or diffusion models (DMs) to model quantized waveform representations or spectral features. Models like MusicLM [Agostinelli *et al.*, 2023], MusicGen [Copet *et al.*, 2024], MeLoDy [Lam *et al.*, 2024], and Jen-1 [Li *et al.*, 2024b] leverage LMs and DMs on residual codebooks obtained via quantization-based codecs [Zeghidour *et al.*, 2021; Défossez *et al.*, 2022]. Moûsai [Schneider *et al.*, 2023], Noise2Music [Huang *et al.*, 2023a], Riffusion [Forsgren and Martiros, 2022], AudioLDM 2 [Liu *et al.*, 2023a], and Stable Audio [Evans *et al.*, 2024a] use U-Net-related diffusion to model mel-spectrograms or latent representations obtained through compression networks. Although some approaches attempt to guide the model towards generating high-quality content by setting negative prompts like "low quality" [Liu *et al.*, 2023a; Chen *et al.*, 2024b], few explicitly inject quality information during training. This results in the model's inability to effectively perceive and control content quality.

Transformer based diffusion models. Traditional diffusion models typically use U-Net as the backbone, where the inductive biases of CNNs do not effectively model the spatial correlations of signals and are insensitive to scaling laws [Li *et al.*, 2024a]. However, transformer-based diffusion models (DiT) [Peebles and Xie, 2023] have effectively addressed these issues. This advantage is particularly evident in fields such as video generation [Brooks *et al.*,], image generation [Peebles and Xie, 2023; Chen *et al.*, 2024a; Bao *et al.*, 2022], and speech generation [Liu *et al.*, 2023c]. To expedite training and foster inter-domain learning of correlations, the masking strategy has proven effective, yielding SOTA class-conditioned performances on ImageNet [Gao *et al.*, 2023]. Additionally, a simpler architecture [Zheng *et al.*, 2023] incorporating reconstruction losses and unmasked finetuning further enhances model training speed. However, these models have not yet been verified for text-controlled music generation on large-scale music datasets, and their adaptability with additional control information remains an open question. Make-an-audio 2 [Huang *et al.*, 2023b] and, more recently, Stable Audio 2 [Evans *et al.*, 2024b], have explored the DiT architecture for audio and sound generation. However, their approach models latent tokens by segmenting only along the time dimension to control and extend generation duration. In contrast, our focus is on finer segmentation within the latent space across both time and frequency, aiming for more precise modeling of music signals.

Quality enhancement in audio domain. Previous research has made efforts to improve the quality of generated audio, particularly in two key areas: waveform fidelity and the consistency between input text and generated content. Waveform quality can be compromised by issues like aliasing from low sampling rates and limited expressiveness due to monophonic representations, while models like MusicGen [Copet *et al.*, 2024] and Stable Audio [Evans *et al.*, 2024a; Evans *et al.*, 2024b], which directly model 32k and 44.1k stereo audio, have significantly enhanced perceptual quality. Despite higher sampling rates and channels, the quality of audio in training datasets remains inconsistent, often suffering from noise, dullness, and a lack of rhythm or structure. These problems, often reflected by the Mean Opinion Score (MOS), are rarely addressed. In terms of text-audio consistency, Make-an-audio 2 [Huang *et al.*, 2023b] and WavCaps [Mei *et al.*, 2024] have employed ChatGPT-assisted data augmentation to improve temporal relationships and accuracy in audio effect generation. Although studies like Music-llama [Liu *et al.*, 2024] and LP-musiccaps [Doh *et al.*, 2023] have introduced captioning approaches for music, few have explored the augmentation and utilization of synthetic data in large-scale music generation tasks.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

3 Preliminary

Latent diffusion model. Direct application of DMs to cope with distributions of raw signals incurs significant computational overhead [Ho *et al.*, 2020; Song *et al.*, 2020]. Conversely, studies [Liu *et al.*, 2023b; Liu *et al.*, 2023a] apply them in a latent space with fewer dimensions. The latent representation z_0 is the ultimate prediction target for DMs, which involve two key processes: diffusion and reverse processes. In the diffusion process, Gaussian noise is incrementally added to the original representation at each time step t, described by $z_{t+1} = \sqrt{1-\beta_t}z_t + \sqrt{\beta_t}\epsilon$, where ϵ is drawn from a standard normal distribution $\mathcal{N}(0, I)$, and β_t is gradually adapted based on a preset schedule to progressively introduce noise into the state z_t. The cost function [Ho *et al.*, 2020; Liu *et al.*, 2023b] is formalized as $\arg\min_\theta \mathbb{E}_{(z_0,y),\epsilon}\left[\left\|\epsilon - D_\theta\left(\sqrt{\alpha_t}z_0 + \sqrt{1-\alpha_t}\epsilon, t, y\right)\right\|^2\right]$. where D_θ, the denoising model, strives to estimate the Gaussian noise ϵ, conditioned on the latent state z_t, the time step t, the conditional embedding y, and where α_t represents a predefined monotonically increasing function. In the reverse process, we obtain z_{t-1} via the recursive equation: $z_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta\right) + \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t}\epsilon$, where ϵ_θ represents the estimated Gaussian noise.

Classifier-free guidance. Classifier-free guidance (CFG), introduced by [Ho *et al.*, 2020], increases the versatility and flexible control ability of DMs by both considering conditional and unconditional generation. Typically, a diffusion model generates content based on specific control signals y within its denoising function $D_\theta(z_t, t, y)$. CFG enhances this mechanism by incorporating an unconditional mode $D_\theta(z_t, t, \emptyset)$, where $\emptyset$ symbolizes the absence of specific control signals. The CFG-enhanced denoising function is then expressed as $D_\theta^{\text{CFG}}(z_t, t, y) = D_\theta(z_t, t, y) + w(D_\theta(z_t, t, y) - D_\theta(z_t, t, \emptyset))$, where $w \geq 1$ denotes the guidance scale. During training, the model substitutes y with $\emptyset$ at a constant probability p_{uncond}. In inference, $\emptyset$ might be replaced by a negative prompt like "low quality" to prevent the model from producing such attributes [Liu *et al.*, 2023a].

4 Method

4.1 Quality Information Injection

At the heart of our work lies the implementation of a pseudo-MOS scoring model [Ragano *et al.*, 2023] to meticulously assign music quality to quality prefixes and quality tokens.

We define our training set as $\mathcal{D}_o = \{(M_i, T_i^o) \mid i = 1, 2, \ldots, N_D\}$, where each M_i represents a music signal and T_i^o is the corresponding original textual description. To optimize model learning from datasets with diverse audio quality and minimize the impact of low-quality audio, we initially assign *p*-MOS scores to each music track using a model fine-tuned with wav2vec 2.0 [Baevski *et al.*, 2020] on a dataset of vinyl recordings for audio quality assessment, and achieve the corresponding *p*-MOS set $S = \{s_1, s_2, \ldots, s_{N_D}\}$. These scores facilitate dual-perspective quality control for enhanced granularity and precision.

First, We analyze this *p*-MOS set S to identify a negative skew normal distribution with mean μ and variance σ^2. We define text prefixes based on s as follows: prepend "low quality" if $s < \mu - 2\sigma$, "medium quality" if $\mu - \sigma \leq m \leq \mu + \sigma$, and "high quality" if $s > \mu + 2\sigma$. This information is prepended before processing through the text encoder with cross-attention, enabling the initial separation of quality-related information.

To achieve a more precise awareness and control of waveform quality, we synergize the role of text control with quality embedding. We observed that the distribution of *p*-MOS in the dataset is approximately normal, which can be shown in Figure 1, allowing us to use the Empirical Rule to segment the data accordingly. Specifically, we define the quantization function $Q : [0, 5] \to \{1, 2, 3, 4, 5\}$ to map the *p*-MOS scores to discrete levels based on the distance from the mean μ in terms of standard deviation σ:

$$Q(s) = \left\lfloor \frac{s - (\mu - 2\sigma)}{\sigma} \right\rfloor + r \tag{1}$$

where $r = 2$ for $s > \mu$, otherwise, $r = 1$. Subsequently, $Q(s)$ is mapped to a d-dimensional quality vector embedding using the embedding function E, such that

$$q_{\text{vq}}(s) = E(Q(s)) \in \mathbb{R}^d, \tag{2}$$

This process provides finer granularity of control within the following model and facilitates the ability of interpolative quality control during inference, enabling precise adjustments in $\mathbb{R}^d$. In later stages, the quality embedding is treated as a token on par with every latent audio patch, participating in the attention computations to enable interaction.

4.2 Quality-aware Masked Diffusion Transformer

In a general patchify phase with patch size $p_f \times p_l$ and overlap size $o_f \times o_l$, patchified token sequence $X = \{x_1, x_2, \ldots, x_P\} \subset \mathbb{R}^{p_f \times p_l}$ are obtained through spliting the music latent space $\mathcal{M}_{spec} \in \mathbb{R}^{F \times L}$, as described in Section 5.2. The total number of patches P is given by:

$$P = \left\lceil \frac{L - p_l}{p_l - o_l} + 1 \right\rceil \times \left\lceil \frac{F - p_f}{p_f - o_f} + 1 \right\rceil \tag{3}$$

A 2D-Rope position embedding [Su *et al.*, 2024] is added to each patch for better modeling of relative position relationship while a binary mask $\boldsymbol{m} \in \{0, 1\}^P$ is applied during the training stage, with a variable mask ratio γ. This results in a subset of $\lfloor \gamma P \rfloor$ patches being masked that $\sum_{i=1}^{P_N} m_i = \lfloor \gamma P \rfloor$, leaving $P - \lfloor \gamma P \rfloor$ patches unmasked. The subset of masked tokens is invisible in the encoder stage and replaced with trainable mask tokens in the decoder stage following the same strategy utilized in AudioMAE [Huang *et al.*, 2022] and MDT [Gao *et al.*, 2023].

The transformer we use consists of N encoder blocks, M decoder blocks, and an intermediate layer to replace the masked part with trainable parameters. We treat the embedding of the quantized *p*-MOS score as a prefix token, concatenated with each stage's music tokens. Let $X^k = [x_1^k, x_2^k, \ldots, x_P^k] \in \mathbb{R}^{P \times d}$ represent the output of k-th encoder or decoder block, where the initial input of the encoder

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

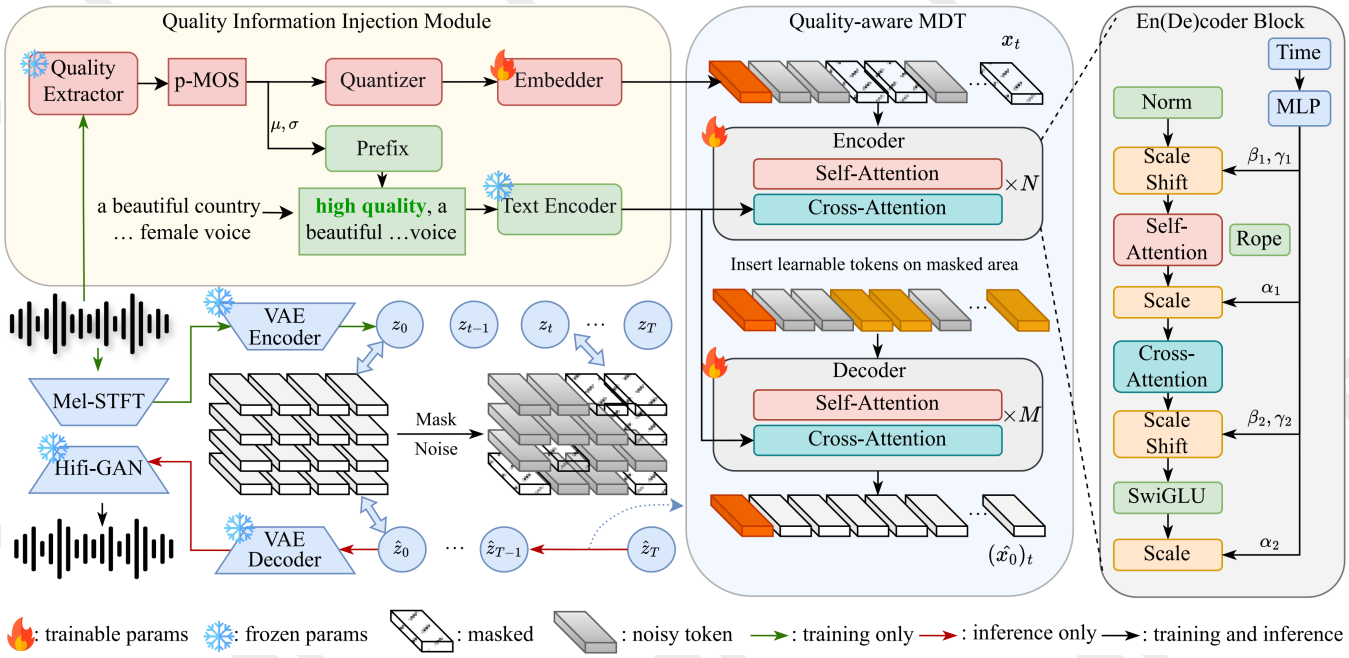

Figure 2: Pipeline of proposed quality-aware masked diffusion transformer for music generation.

$X^0 = z_t = \alpha_t z_0 + \sqrt{1-\alpha_t}\epsilon$, and the final decoder block estimate $X^{N+M} = z_0 = [x_1, x_2, \ldots, x_P]$. For $k < N$, indicating the encoder blocks, the sequence transformation focuses only on unmasked tokens:

$$[q_{\text{vq}}^{k+1}; X^{k+1}] = \text{Encoder}^k\left([q_{\text{vq}}; X^k \odot (\mathbf{1} - \boldsymbol{m})]\right), \quad (4)$$

where $\boldsymbol{m} \in \{0,1\}^P$ is the mask vector, with 1 indicating masked positions and 0 for visible tokens.

For $N < k < N + M$, indicating the decoder blocks, the full sequence including both unmasked tokens and learnable masked tokens is considered:

$$[q_{\text{vq}}^{k+1}; X^{k+1}] = \text{Decoder}^k\left([q_{\text{vq}}; X^k]\right), \quad (5)$$

where the previously masked tokens are now subject to prediction and refinement. In the decoding phase, the portions that were masked are gradually predicted, and throughout this entire phase, the quality token $q_{\text{vq}}(s)$ is progressively infused and optimized. Subsequently, the split patches are unpatchified while the overlapped area is averaged to reconstruct the output noise and every token contributes to calculating the final loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{(z_0, q_{vq}, y), \epsilon}\left[\left\|\epsilon - D_\theta\left(\sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon, t, q_{vq}, y\right)\right\|^2\right] \quad (6)$$

In the inference stage, the model can be guided to generate high-quality music through modified CFG:

$$D_\theta^{\text{High}}(z_t, t, q_{vq}^{\text{high}}, y) = D_\theta(z_t, t, q_{vq}^{\text{high}}, y) + w\left(D_\theta(z_t, t, q_{vq}^{\text{high}}, y) - D_\theta(z_t, t, q_{vq}^{\text{low}}, \emptyset)\right) \quad (7)$$

Here q_{vq}^{high} and q_{vq}^{low} indicate quantified p-MOS for guiding the model in a balance between generation quality and diversity. After obtaining the sampled latent $\hat{z}_0$ with DDIM sampler, we can finally generate the music using the VAE decoder.

4.3 Music Caption Refinement

We divided the caption refinement stage into three steps including text information enriching with music caption model $\mathcal{F}_{\text{cap}}$, caption adjustment with CLAP cosine similarity function $\mathcal{S}$ and caption diversity extension with LLMs which we denoted as $\mathcal{F}_{\text{llm}}$.

Initially, pretrained music caption model [Doh *et al.*, 2023] is employed to re-annotate each music signal M_i to T_i^g, shown as $\mathcal{D}_g = \{(M_i, T_i^g) \mid T_i^g = \mathcal{F}_{\text{cap}}(M_i), i = 1, 2, \ldots, N\}$. CLAP text-audio similarity is applied to filter $\mathcal{D}_g$ with a threshold of ρ_1, resulting in

$$\mathcal{D}_g^{\text{filter}} = \{(M_i, T_i^g) \mid \mathcal{S}(T_i^g, M_i) > \rho_1\} \quad (8)$$

In this context, we meticulously filter out generated captions that do not correspond with their respective audio files. This misalignment may be attributed to inaccuracies within the captioner's insufficient training. For the filtered data pairs, we opt to retain the use of the original captions.

To ensure that valuable information from the original captions is not overlooked when using only the generated captions, we adapt a fusing stage to combine the original caption and generated pseudo prompt. Firstly, we need to filter out original captions that is useless or inaccurate, formulated as:

$$\mathcal{D}_o^{\text{filter}} = \{(M_i, T_i^o) \mid \mathcal{S}(T_i^o, M_i) > \rho_2\}. \quad (9)$$

The issue can stem from the original data being improperly labeled with terms such as 'speech, car' from datasets like AudioSet [Gemmeke *et al.*, 2017] and also may be because of desperately missing of the original labels.

Finally, only the original caption that suffers low CLAP text similarity score should be merged with the generated ones, for redundant, repetitive parts result in long and verbose

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

final captions. Thus, we set the threshold to ρ_3 and merge them by LLMs to $T_{\text{fusion}} = \mathcal{F}_{\text{llm}}(T^o, T^g)$:

$$\mathcal{D}_{\text{merge}} = \big\{(M_i, T_{\text{fusion}}) \mid \mathcal{S}(T^o, T^g) < \rho_3, \\ (M_i, T^o) \in \mathcal{D}_o^{\text{filter}}, (M_i, T^g) \in \mathcal{D}_g^{\text{filter}}\big\}. \tag{10}$$

5 Experimental Setup

5.1 Datasets

For training, we used the following databases for our training: AudioSet Music Subset (ASM) [Gemmeke *et al.*, 2017], MagnaTagTune (MTT) [Law *et al.*, 2009], Million Song Dataset (MSD) [Bertin-Mahieux *et al.*, 2011], Free Music Archive (FMA) [Defferrard *et al.*, 2016], and an additional dataset[1]. Each track in these databases was clipped to 10-second segments and sampled at 16kHz to ensure uniformity across the dataset. The final training set was developed through a process of caption refinement, as detailed in Section 4.3. Finally, we got our training set totaling 12.5k hours of diverse music data. The specific composition of these datasets is further elaborated in the Appendix. For evaluation, we test our model on the widely used MusicCaps benchmark [Agostinelli *et al.*, 2023] and the Song-Describer-Dataset [Manco *et al.*, 2023]. MusicCaps consists of 5.5K 10.24-second clips sourced from YouTube, each accompanied by high-quality music descriptions provided by ten musicians. The Song-Describer Dataset is made up of 706 licensed high quality music recordings.

5.2 Models and Hyperparameters

Audio compression. Each 10.24-second audio clip, sampled at 16 kHz, is initially transformed into a 64×1024 mel-spectrogram with mel-bins of 64, hop-length of 160 and window length of 1024. Subsequently, this spectrogram is compressed into a 16×128 latent representation $\mathcal{M}_{spec}$ using a Variational Autoencoder (VAE) pretrained with AudioLDM 2 [Liu *et al.*, 2023a] with series of quantization loss and adversarial loss. We use pretrained Hifi-GAN [Kong *et al.*, 2020] vocoder to reconstruct the waveform from the generated mel-spectrogram.

Caption processing and conditioning. We utilize the LP-MusicCaps [Doh *et al.*, 2023] caption model for ASM, FMA, and subsets of MTT and MSD that have weak or no captions. We use the official checkpoint from LAION-CLAP [Wu *et al.*, 2023][2] for text-to-text and text-to-audio similarity calculations. Based on small scale subjective experienment, thresholds are set at $\rho_1 = \rho_2 = 0.1$ to ensure any generated text or original caption not aligned well with the corresponding waveform is filtered out. Additionally, after filtering, generated text that fall below a threshold of $\rho_3 = 0.25$ are merged with original tags with the prompt: *Merge this music caption "generated caption" with the ground truth tags "original tags", and do not add any imaginary elements.*. We use FLAN-T5-large [Peebles and Xie, 2023] as text encoder for all models.

[1]We use 55k music tracks from https://pixabay.com, which is a large scale copyright free dataset.

[2]`music_speech_audioset_epoch_15_esc_89.98.pt`

Diffusion backbone. We train our diffusion model with three backbones for comparison: U-Net [Ronneberger *et al.*, 2015] based at 1.0B parameters; our proposed Quality-aware Masked Diffusion Transformer (QA-MDT) with $N = 20$ encoder layers and $M = 8$ decoder layers at 675M. We study the impact of the patch size and overlap size in the Appendix, and apply a patch size of 1×4 without overlap for the training of our final model. We train on 10.24-second audio crops sampled at random from the full track, maintaining a total batch size of 64, learning rate of 8e-5, and a condition drop of 0.1 during training. The final model was trained for a total of 38.5k steps. During inference, we use Denoising Diffusion Implicit Models (DDIM) [Song *et al.*, 2020] with 200 steps and a guidance scale of 3.5, consistent with AudioLDM [Liu *et al.*, 2023b]. We begin by presenting our approach to refining captioning, which includes the capability for quality awareness, transitioning from text-level control to token-level control. Finally, we compare proposed model with previous works subjectively and objectively.

5.3 Evaluation Metrics

We evaluate the proposed method using objective metrics, including the Fréchet Audio Distance (FAD) [Kilgour *et al.*, 2018], Kullback-Leibler Divergence (KL), Inception Score (IS)[3]. We also utilize pseudo-MOS scoring model [Ragano *et al.*, 2023] to estimate generation quality, with more accurate assessments derived from subjective metrics.

6 Results

6.1 Quality Awareness

This subsection explores the effects and interactions of model control over quality tokens and quality text prefixes during the training phase, as well as their comparative effects across different models. In our previous MTT dataset of 1,000 test pairs, we filtered out pairs labeled with *low quality* or *quality is poor* to avoid confusion when applying quality prefixes, resulting in a new subset of 519 entries, which we refer to as the MTT Filter Set (MTT-FS). Figure 4 illustrates the impact of different quality prefixes during inference when quality is used as a text prefix during training for U-Net and MDT-based backbones. It was observed that U-Net, when inferred with different quality prefixes, showed only minor changes in *p*-MOS scores and did not adhere to the threshold set during training. In contrast, MDT demonstrated better learning of quality information from prefixes, achieving *p*-MOS scores significantly higher than those of U-Net and the test set. Additionally, by decoupling quality information from the training set, we achieved better FAD (5.602 vs 5.757) and higher *p*-MOS (4.039 vs 3.796) compared to training and inference without quality text prefixes. Given that quality tokens are specifically designed for the Transformer architecture, Figure 3 (left) shows the controlled outcomes

[3]We **strictly follow** the comparison method and evalution code in AudioLDM2 and ensure that the indicators in the paper **follow consistent sampling rates and durations for fair comparison**. All above metrics are computed using the `audioldm_eval` library [Liu *et al.*, 2023b], ensuring standardized evaluation.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

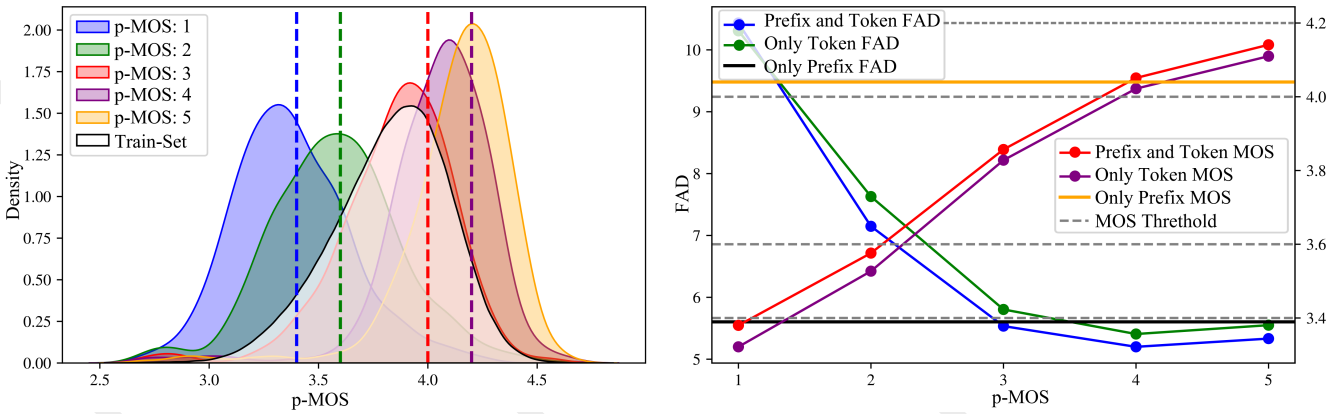

Figure 3: **(Left)** Five p-MOS distribution curves are obtained by concurrently using text quality prefixes and quality tokens as controls on the MTT-FS, with quantized MOS levels ranging from 1 to 5 serving as control constraint inferences. The distribution of the training set is normalized by each sample's duration, colored lines represent thresholds of quantized p-MOS tokens during training. **(Right)** The effect of using quality text prefixes during training is shown, showcasing testing results on FAD and p-MOS, while gray lines for quantized p-MOS threshold.

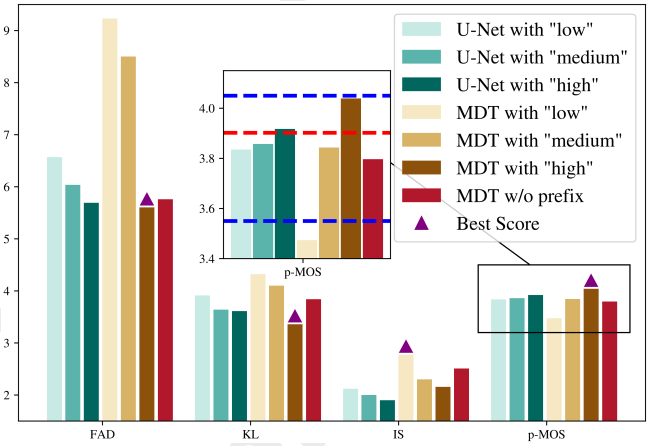

Figure 4: Comparison of model performance under different quality prefixes on MTT-FS, while the blue dashed line represents the threshold set during training to distinguish the three quality prefix levels, and the red one represents the test set average p-MOS value.

when different quality tokens are used after integrating quantified quality as a token during training. Remarkably, using quality tokens alone provided more precise and accurate p-MOS score control. In our ablation study, we compared the effects of using only text prefixes against combining both approaches. As shown in Figure 3 (right), as the quantized control level gradually increased, the model steadily improved in p-MOS scores, which represent the quality of generation. Concurrently, FAD and KL also progressively optimized until a turning point at level 4, where a higher average p-MOS was achieved than when solely using prefixes. This turning point may be due to the scarcity of examples with quality level 5 in the dataset. Moreover, by combining two types of quality information injection, the refined decoupling and interaction allowed the model to more accurately perceive audio data quality features during training, leading to significant reductions in FAD and KL compared to using only one of them.

We also compare our approach with the traditional "negative prompt" strategy [Liu *et al.*, 2023a] in Appendix, highlighting our approach's significant improvement in quality and reduction in FAD.

6.2 Impact of Music Caption Refinement

	U-Net based			MDT based		
Caption	FAD ↓	IS↑	CLAP ↑	FAD ↓	IS ↑	CLAP ↑
$\mathcal{D}_o$	7.23	1.74	0.199	7.07	2.12	0.291
$\mathcal{D}_g$	5.94	2.28	0.278	5.76	2.51	0.342
$\mathcal{D}_{\text{merge}}$	**5.87**	**2.29**	**0.284**	**5.64**	**2.63**	**0.350**

Table 1: Comparison of model performance training on different textual representations, evaluated by FAD, IS and CLAP score.

We conducted our ablation study on a subset of our training set, which includes ASM and FMA, totaling approximately 3,700 hours and 1.1 million clips. For evaluation, we utilized an out-of-domain set with 1,000 samples randomly selected from MTT [Law *et al.*, 2009]. Table 2 compares the model's performance using different textual representations: sentences formed by merging original tags with commas ($\mathcal{D}_o$), generated captions ($\mathcal{D}_g$), and generated captions refined through filtering and fusion ($\mathcal{D}_{\text{merge}}$). During the filtering and fusion stage, 8.9% of the generated captions were filtered out, and 15.1% were fused with original tags using ChatGPT. Each model underwent training for 60,000 steps with a batch size of 64.

From Table 2 we can also observe consistent trends: employing a captioner to transform audio annotations from sparse words into detailed sentences significantly improved the models' generalization and diversity. This indicates that detailed annotations are essential for learning the relationship between the models and spectral features. Moreover, the fil-

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

	Details		MusicCaps				Song Describer Dataset			
Model	Params	Hours	FAD ↓	KL ↓	IS ↑	CLAP ↑	FAD ↓	KL ↓	IS ↑	CLAP ↑
MusicLM	1290M	280k	4.00	–	–	–	–	–	–	–
MusicGen†	1.5B	20k	3.80	1.22	–	0.31	5.38	1.01	1.92	0.18
Mousai	1042M	2.5k	7.50	1.59	–	0.23	–	–	–	–
Jen-1	746M	5.0k	2.00	1.29	–	0.33	–	–	–	–
AudioLDM 2 – Full	712M	17.9k	3.13	**1.20**	–	–	–	–	–	–
AudioLDM 2 – Music†	712M	10.8k	4.04	1.46	2.67	0.34	2.77	0.84	1.91	0.28
Ours (U-Net)	1.0B	12.5k	2.03	1.51	2.41	0.33	**1.01**	**0.83**	1.92	0.30
Ours (QA-MDT)	**675M**	12.5k	**1.65**	1.31	**2.80**	**0.35**	1.04	**0.83**	**1.94**	**0.32**

Table 2: Objective evaluation results for music generation with diffusion-based and language-model-based approaches. Methods we re-inferred are marked with †.

	Po		Pmp		Ve		Bg	
Model	Ovl	Rel	Ovl	Rel	Ovl	Rel	Ovl	Rel
Ground Truth	4.00	4.00	4.47	3.60	4.10	3.80	3.87	3.87
AudioLDM 2	2.03	2.42	3.03	3.61	3.21	3.71	3.85	3.85
MusicGen	2.83	3.54	2.63	2.92	3.41	3.00	**4.33**	3.83
Ours(U-Net)	2.80	3.34	3.46	4.08	3.40	**3.96**	3.88	3.96
Ours(QA-MDT)	**3.27**	**3.77**	**3.69**	**4.19**	**3.54**	3.94	4.23	**4.00**

Table 3: Evaluation of model performances among different groups, rated for text relevance (Rel) and overall quality (Ovl), with higher scores indicating better performance. The groups included Production Operators (Po), Professional Music Producers (Pmp), Video Editors (Ve) and Beginners(Bg)

ter and fusion stages led to enhancements across all metrics, highlighting the significance of precise, comprehensive annotations for generalization ability and control ability. We also found that compared to U-Net, the MDT architecture shows stable improvements in basic modeling metrics, making it a better backbone for music spectral modeling.

6.3 Compared with Previous Methods

We compared our proposed method with the following representative previous methods: AudioLDM 2 [Liu *et al.*, 2023a], Mousai [Schneider *et al.*, 2023] and Jen-1 [Li *et al.*, 2023] which model music using spectral latent spaces, MusicLM [Agostinelli *et al.*, 2023], and MusicGen [Copet *et al.*, 2024], which focus on modeling discrete representations.

We re-inferred AudioLDM2-Music and MusicGen-1.5B using their official checkpoints to compare additional metrics under the same environment. The results are presented in Table 2. For Ours (U-Net), we inferred all text with the prefix "high quality", while for Ours (QA-MDT), we used the same prefix along with a *p*-MOS quality token set to level 5. When calculating the CLAP score, we evaluated the generated music with original prompt, which did not include any quality prefix. The experimental results show significant advantages in both subjective and objective metrics for our models. Since KL divergence measures the distance between audio samples, higher quality audio often results in deviations from the original waveform of Musiccaps, which can lead to lower performance. Although Ours (U-Net) showed a slight FAD advantage on the Song-Describer-Dataset, this may be due to instabilities arising from the small scale test dataset, and we further demonstrated the superiority of QA-MDT in subsequent subjective experiments. Additionally, since MusicGen was trained on non-vocal tracks, it may underperform on captions that include vocals.

Based on subjective evaluation shown in Table 3, our proposed method significantly improves overall audio quality and text-audio consistency, thanks to the label optimization for large music datasets and the quality-aware training strategy. By analyzing the backgrounds of the evaluators and their corresponding results, we can also see that for beginners, the comparison between different systems is not sensitive, which is related to their lack of music background experience and knowledge. However, from the perspective of our method in product operators, video editors, and audio producers, our method offers considerable enhancements, underscoring its potential value to audio industry professionals.

7 Conclusion and Discussion

In this study, we address the key challenges in the music generation domain, including model architecture design, large-scale uneven audio quality, and unaligned textual annotations, all of which impede the progress of TTM with quality, musicality, and text alignment. In the future, we aim to further enhance and expand our model to achieve long-duration, high-sampling-rate, controllable, and highly interactive music generation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62171427.

Contribution statement

Chang Li and Ruoyu Wang contributed equally. Jun Du supervised the entire project and serves as the corresponding author.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

References

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

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[Bao *et al.*, 2022] Fan Bao, Chongxuan Li, Yue Cao, and Jun Zhu. All are worth words: a vit backbone for score-based diffusion models. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[Bertin-Mahieux *et al.*, 2011] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. 2011.

[Brooks *et al.*,] Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators, 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*.

[Chen *et al.*, 2024a] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024.

[Chen *et al.*, 2024b] Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1206–1210. IEEE, 2024.

[Copet *et al.*, 2024] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[Defferrard *et al.*, 2016] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

[Défossez *et al.*, 2022] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[Doh *et al.*, 2023] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023.

[Evans *et al.*, 2024a] Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*, 2024.

[Evans *et al.*, 2024b] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*, 2024.

[Forsgren and Martiros, 2022] Seth* Forsgren and Hayk* Martiros. Riffusion - Stable diffusion for real-time music generation. 2022.

[Gao *et al.*, 2023] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23164–23173, 2023.

[Gemmeke *et al.*, 2017] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Huang *et al.*, 2022] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.

[Huang *et al.*, 2023a] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023.

[Huang *et al.*, 2023b] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.

[Kilgour *et al.*, 2018] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

[Kong *et al.*, 2020] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[Lam *et al.*, 2024] Max WY Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, et al. Efficient neural music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[Law *et al.*, 2009] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

using games: The case of music tagging. In *ISMIR*, pages 387–392. Citeseer, 2009.

[Li *et al.*, 2023] Peike Li, Boyu Chen, Yao Yao, Yikai Wang, Allen Wang, and Alex Wang. Jen-1: Text-guided universal music generation with omnidirectional diffusion models. *arXiv preprint arXiv:2308.04729*, 2023.

[Li *et al.*, 2024a] Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R Manmatha, Ashwin Swaminathan, Zhuowen Tu, Stefano Ermon, and Stefano Soatto. On the scalability of diffusion-based text-to-image generation. *arXiv preprint arXiv:2404.02883*, 2024.

[Li *et al.*, 2024b] Peike Patrick Li, Boyu Chen, Yao Yao, Yikai Wang, Allen Wang, and Alex Wang. Jen-1: Text-guided universal music generation with omnidirectional diffusion models. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 762–769. IEEE, 2024.

[Liu *et al.*, 2023a] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023.

[Liu *et al.*, 2023b] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023.

[Liu *et al.*, 2023c] Huadai Liu, Rongjie Huang, Xuan Lin, Wenqiang Xu, Maozong Zheng, Hong Chen, Jinzheng He, and Zhou Zhao. Vit-tts: visual text-to-speech with scalable diffusion transformer. *arXiv preprint arXiv:2305.12708*, 2023.

[Liu *et al.*, 2024] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE, 2024.

[Manco *et al.*, 2023] Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: a corpus of audio captions for music-and-language evaluation. In *Machine Learning for Audio Workshop at NeurIPS 2023*, 2023.

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

[Peebles and Xie, 2023] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[Ragano *et al.*, 2023] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines. Audio quality assessment of vinyl music collections using self-supervised learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Rombach *et al.*, 2021] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[Schneider *et al.*, 2023] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Mo\^usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Su *et al.*, 2024] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[Wu *et al.*, 2023] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[Zeghidour *et al.*, 2021] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

[Zheng *et al.*, 2023] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.