Generative Music Models' Alignment with Professional and Amateur Users' Expectations

Zihao Wang[1,3], Jiaxing Yu[1], Haoxuan Liu[1], Zehui Zheng[1]
Yuhang Jin[1], Shuyu Li[1], Shulei Ji[1,2], Kejun Zhang[1,2*]

[1]Zhejiang University
[2]Innovation Center of Yangtze River Delta, Zhejiang University
[3]Carnegie Mellon University

[*]Corresponding Author: zhangkejun@zju.edu.cn

Abstract

Recent years have witnessed rapid advancements in text-to-music generation using large language models, yielding notable outputs. A critical challenge is understanding users with diverse musical expertise and generating music that meets their expectations, an area that remains underexplored. To address this gap, we introduce the novel task of Professional and Amateur Description-to-Song Generation. This task focuses on aligning generated content with human expressions from varying musical proficiency levels, aiming to produce songs that accurately meet auditory expectations and adhere to musical structural conventions. We utilized the MuChin dataset, which contains annotations from both professionals and amateurs for identical songs, as the source for these distinct description types. We also collected a pre-train dataset of over 1.5 million songs; lyrics were included for some, while for others, lyrics were generated using Automatic Speech Recognition (ASR) models. Furthermore, we propose MuDiT/MuSiT, a single-stage framework designed to enhance human-machine alignment in song generation. This framework employs Chinese MuLan (ChinMu) for cross-modal comprehension between natural language descriptions and auditory musical attributes, thereby aligning generated songs with user-defined outcomes. Concurrently, a DiT/SiT model facilitates end-to-end generation of complete songs audio, encompassing both vocals and instrumentation. We proposed metrics to evaluate semantic and auditory discrepancies between generated content and target music. Experimental results demonstrate that MuDiT/MuSiT outperforms baseline models and exhibits superior alignment with both professional and amateur song descriptions.

1 Introduction

Music is a universal language that transcends cultural barriers, yet the automated creation of music that aligns with human thoughts is a complex endeavor. In recent years, large language models for text-to-music generation have advanced rapidly, resulting in notable systems. A key challenge for these models is aligning generated music with the diverse expectations of users with varying musical expertise. However, this issue remains underexplored.

To address this gap, we introduce the "Professional and Amateur Description-to-Song Generation" task, which prioritizes the alignment of AI-generated music with both professional and amateur human expressions. We propose MuDiT/-MuSiT, a single-stage framework based on two commonly used diffusion-transformer-like models (Peebles and Xie, 2023; Ma et al., 2024) for music generation. To address the limited coverage of Chinese colloquial phrases in existing text-to-audio pre-trained models, we trained a Chinese MuLan (ChinMu) cross-modal encoder based on CLAP (Wu et al., 2023b) and MuLan (Huang et al., 2022) architectures. The MuDiT/MuSiT framework integrates cross-modal understanding via ChinMu to generate music that aligns with user auditory expectations and musical conventions. During inference, ChinMu converts user colloquial descriptions into vectors. These vectors, concatenated with random noise, serve as conditional inputs to DiT/SiT modules. Operating in the VAE (Kingma and Welling, 2013) latent space, DiT and SiT then generate songs corresponding to these descriptions. LLM-generated lyrics can also be incorporated as conditions via a cross-attention mechanism.

MuDiT/MuSiT underwent a three-stage training process: module preparation, large-scale data pretraining, and annotated data fine-tuning. For finetuning, we utilized the MuChin dataset (Wang et al., 2024b). Unlike datasets often limited to expert-only annotations or those with semantic gaps from automated Music Information Retrieval (MIR) tag-

Findings of the Association for Computational Linguistics: ACL 2025, pages 6909–6920
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ging, MuChin provides paired professional and amateur annotations for identical songs. This unique characteristic, previously validated for its richness with models like MERT (Li et al., 2023) and Jukebox (Dhariwal et al., 2020), makes it highly suitable for training our model to align with these varied user expressions.

Models were evaluated using subjective and objective metrics, focusing particularly on their ability to understand meanings conveyed by users with varying musical expertise (amateur or professional) and to generate music meeting user expectations. In addition to Fréchet Audio Distance (FAD) for assessing generation quality, we introduced SST and ASOA—novel metrics for semantic and auditory alignment with input descriptions. Given that music generation lacks standard references, unlike text annotation, we supplemented with human evaluation as subjective metrics.

For experimental comparison of alignment performance, we selected AudioLDM (Liu et al., 2023a), StableAudio (Evans et al., 2024), and MusicGen (Copet et al., 2024) as baseline models. Results indicate that MuDiT/MuSiT surpasses baseline models in aligning with professional and amateur colloquial expressions, marking a significant step in human-AI collaborative music creation. This research, by introducing the "Professional and Amateur Description-to-Song Generation" task, is the first to explicitly address the challenge of generating music that resonates with the everyday language of the general public. It tackles a critical gap in aligning AI-generated music with non-expert colloquial expressions and paves the way for more nuanced human-AI interaction in artistic expression.

Code implementation[1].

2 Background

2.1 Text-to-Song Generation

In recent years, advancements in artificial intelligence have propelled the development of automatic music composition (Sheng et al., 2021; Hsiao et al., 2021; Mao et al., 2023; Yu et al., 2022; Wang et al., 2024a; Agostinelli et al., 2023; Wu et al., 2023a; Wang et al., 2024c; Copet et al., 2024), but the alignment [2] of AI-generated music with the everyday expressions of non-professionals remains largely underexplored. This oversight is significant, as it impacts the ability of AI to understand and meet the diverse musical expectations of the general public.

Early research primarily focused on specific aspects of song generation, such as text-to-instrumental music (Schneider et al., 2023; Huang et al., 2023; Agostinelli et al., 2023; Copet et al., 2024), lyrics-to-melody (Sheng et al., 2021; Yu et al., 2024b), and music score-to-song generation (Zhiqing et al., 2024). However, these models are unable to generate complete songs that include both vocals and accompaniment in a single stage. For descriptive text-to-instrumental music generation, models like MusicLM (Agostinelli et al., 2023) and MusicGen (Copet et al., 2024) use quantization-based audio codecs (Zeghidour et al., 2021; Défossez et al., 2022) to obtain residual codebooks and utilize large language models to generate high-quality audio music. Whereas AudioLDM v1 (Liu et al., 2023a) and v2 (Liu et al., 2024) rely on latent diffusion models, focusing on modeling in latent space for generation. For music score and descriptive text-to-song generation, Melodist (Zhiqing et al., 2024) adopts a two-stage approach, first generating vocals based on melody, lyrics, and descriptive text, and then creating accompaniment based on the vocals and descriptive text.

Recent advancements have made it possible to generate complete songs including both vocals and accompaniment (Lei et al., 2024; Yuan et al., 2025). Text-to-song generation platforms from industry, such as Suno [3] (Yu et al., 2024a), SkyMusic [4], Udio [5], Stable Audio [6], and SeedMusic (Bai et al., 2024) have received widespread attention for their powerful capabilities.

2.2 Transformer-Based Diffusion Models

Traditional diffusion models typically employ U-Net architectures, which are limited by the inductive biases of Convolutional Neural Networks (CNNs), making it difficult to effectively model the spatial correlations of signals and are not sensitive to scaling laws (Li et al., 2024). However, Transformer-based diffusion models (DiT) (Peebles and Xie, 2023) have successfully overcome these limitations and have demonstrated significant

[1] https://github.com/CarlWangChina/MuDiT-MuSiT

[2] https://en.wikipedia.org/wiki/AI_alignment

[3] https://www.suno.ai

[4] https://www.tiangong.cn

[5] https://www.udio.com

[6] https://www.stableaudio.com

advantages in areas such as speech generation (Liu et al., 2023b), image generation (Bao et al., 2022; Peebles and Xie, 2023; Chen et al., 2024), and video generation (Brooks et al., 2024). Meanwhile, Scalable Interpolant Transformers (SiT) (Ma et al., 2024), built upon a DiT backbone, use a flexible interpolation framework to connect two distributions more effectively than standard diffusion models, achieving significant results in terms of efficiency and performance.

2.3 Music Datasets and Benchmarks with Descriptive Annotations

In recent years, music datasets have developed rapidly (Bertin-Mahieux et al., 2011; Bogdanov et al., 2019; Wang et al., 2022; Melechovsky et al., 2023; Lu et al., 2023; Schneider et al., 2023; Zhu et al., 2023). AudioSparx [7] and Mustango (Melechovsky et al., 2023) apply Music Information Retrieval (MIR) tools to extract features such as instruments, style, vocals, and BPM from symbolic music or audio music, and utilize large language models to integrate these features into textual descriptions. MuLaMCap (Huang et al., 2023) uses language models to generate descriptive texts and utilizes MuLan (Huang et al., 2022) to match these texts with music in the datasets. However, a huge semantic gap still exists between descriptive datasets based on automatic annotation and complex human descriptions.

MusicLM (Agostinelli et al., 2023) provides a dataset named MusicCaps, which contains music descriptions annotated by professional musicians. Existing manually annotated datasets are typically limited to annotations by professional musicians and a restricted range of descriptions, which significantly differs from the amateur descriptions provided by the general public (Amer et al., 2013; Mikutta et al., 2014).

Furthermore, although frameworks like AIR-Bench (Yang et al., 2024) and Stable Audio Metrics provide multi-dimensional evaluation, they primarily assess audio quality rather than the consistency between the generated music and user expectations. Recent research, such as MusicEval (Liu et al., 2025), has introduced datasets with expert ratings for generative evaluation, but overlooks the differences in listeners' varying musical proficiency levels, which may affect their expectations.

3 Method

This study adopts supervised learning techniques, complemented by the dual perspectives of professional and amateur annotators, and constructs the training process using an end-to-end single-stage approach. This ensures that the AI-generated songs align with human expectations, while also enabling the AI model to understand and incorporate the standard structure of human songs.

3.1 Training Process of MuDiT/MuSiT

As shown in **Figure 1**, the training process of MuDiT/MuSiT includes three stages. In the first stage, we prepare three modules for MuDiT/MuSiT training: a fine-tuned lyric large language model (LLM), a ChinMu cross-modal encoder, and a VAE (encoder and decoder). In the second stage, within the MuDiT/MuSiT framework, we pre-train DiT/SiT using raw lyrics and audio from MuChin and an additional 1.5 million songs with untagged descriptions. In the third stage, we fine-tune DiT/SiT using a combination of structured lyrics, audio, and professional and amateur descriptions from the MuChin dataset.

3.1.1 Chinese MuLan (ChinMu) Cross-Modal Encoder.

Text-audio contrastive pre-training models are crucial for AI to understand colloquial descriptions and cannot be replaced by other text encoders trained on professional vocabulary. To address the insufficient coverage of Chinese colloquial phrases in existing text-audio contrastive pre-training models, we trained a Chinese MuLan (ChinMu) cross-modal encoder using MuChin, based on the architectures of CLAP (Wu et al., 2023b) and MuLan (Huang et al., 2022). Subsequently, we utilize ChinMu to process the colloquial description into a vector and concatenate it with random noise before inputting it into DiT/SiT as a condition, as shown in **Figure 1**. This step ensures that the input text description is transformed into a dense vector representation capable of capturing semantic nuances. During the training of the ChinMu model, we segmented descriptions and audio of varying lengths from MuChin, enabling effective text-audio contrastive pre-training. This allows ChinMu to adapt to inputs with different text and audio lengths.

[7] https://www.audiosparx.com/

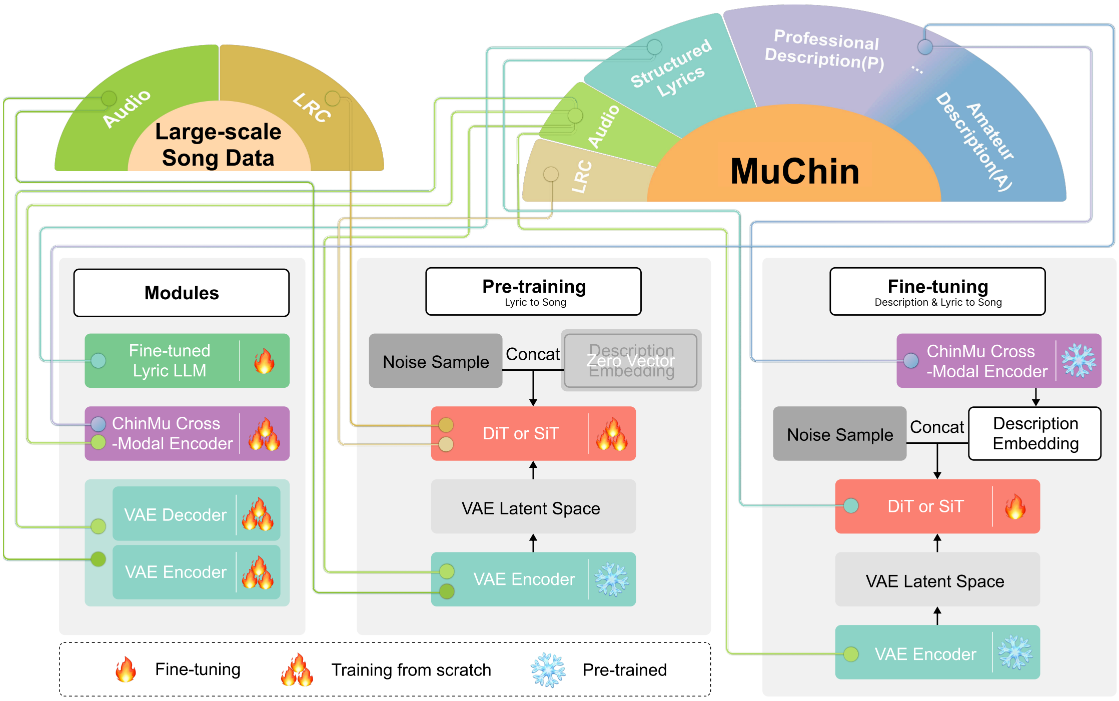


Figure 1: Training details of MuDiT/MuSiT. It includes three stages: module preparation, pre-training, and fine-tuning. In the module preparation stage, we fine-tune a lyric large language model (LLM) using structured lyrics, train a ChinMu cross-modal encoder using audio along with professional and amateur descriptions, and train VAE encoder and decoder using audio. In the pre-training stage, we utilize raw lyrics and audio to pre-train DiT or SiT. During the fine-tuning stage, we optimize the pre-trained DiT or SiT, treating professional and amateur descriptions as well as structured lyrics as conditioning factors.

3.1.2 Fine-tuned Lyric Large Language Model (LLM).

We utilize QLoRA (Dettmers et al., 2024) as a parameter-efficient fine-tuning (PEFT) method to optimize the Qwen-14B model (Bai et al., 2023) for generating lyrics based on colloquial descriptions, incorporating musical sections and rhyming structures. Due to parameter constraints, and considering its effectiveness in Chinese lyric processing, we selected the Qwen-14B-Chat-Int4 model (Bai et al., 2023). Our training data includes themes extracted from lyrics, along with manually annotated musical sections and rhyming structures. This dataset is used to fine-tune the model to generate lyrics that include musical structures (as shown in **Figure 1**), with outputs including tags such as <verse>, <chorus>, and <bridge>. The DiT model generates songs consistent with these tagged structures. Additionally, we convert Chinese characters into Pinyin as input for the DiT model.

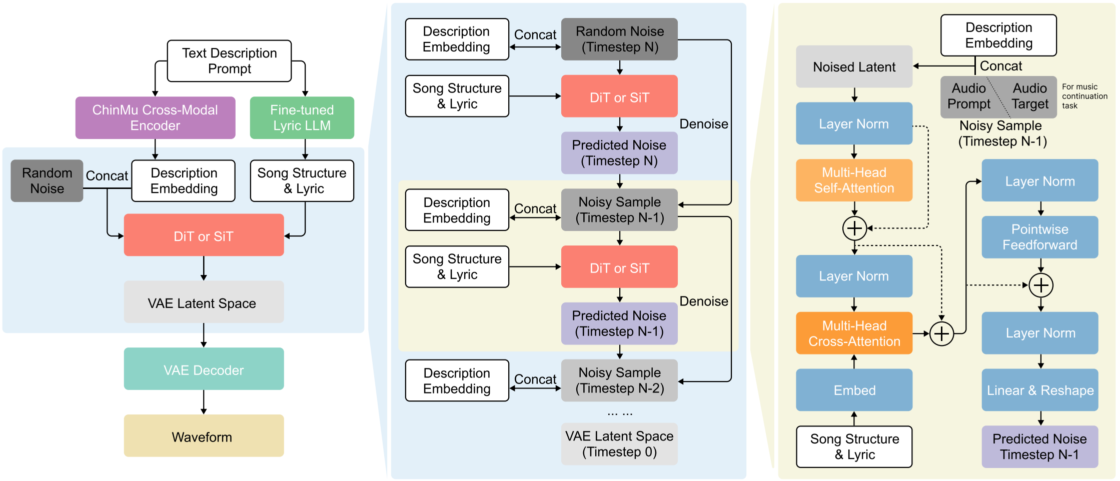
3.1.3 DiT/SiT.

During the training process of DiT/SiT, we employ DDPM with random timesteps, while during inference, we use DDIM with sequential timesteps (progressing from t to 0).

1) Pre-training Phase. We conducted supervised pre-training on DiT/SiT using a collection of 1.5 million de-duplicated songs, as shown in **Figure 1**. An internal automatic speech recognition (ASR) tool extracts lyric texts from the song audio, creating a dataset of "lyric text-song audio" pairs. The lyric text serves as a supervision signal through a cross-attention mechanism, while the song audio, in the form of VAE latent vectors, serves as training data.

During training, lyric timestamps can align audio windows with corresponding lyrics. However, during inference, lyrics provided by users or LLMs often lack precise timestamps, making it difficult for the model to accurately match lyric lengths with audio windows. Compared to speech, the variability in singing speed exacerbates this challenge, complicating the prediction of time ranges based on word counts.

Therefore, we chose not to use common window-based audio generation methods. Instead, we generate the entire song at once, using the full audio without random segmentation during training. To accommodate variable lengths, we add padding to



(a) Overall architecture. (b) Complete diffusion process. (c) One timestep in diffusion process.

Figure 2: Detailed architecture of MuDiT/MuSiT. (a) MuDiT/MuSiT uses ChinMu to convert text into vectors, combines them with noise for DiT input, and uses a fine-tuned LLM to generate lyrics. It outputs complete songs in the form of VAE latent variables, which are then decoded into WAV files. (b) The diffusion process of DiT/SiT starts from noise at t=N, iteratively subtracting the predicted noise from the sample until t=0, outputting the final song as VAE latent variables. (c) In each diffusion timestep, DiT/SiT denoises the sample, processing variable-length lyrics and structures via cross-attention, while fixed-length description embeddings are processed via self-attention; finally outputting noise predictions.

the end of each audio segment for standardization.

This approach resolves the problem of segmenting lyrics into windows and offers two additional benefits: 1) Compared to window-based generation, it enables the DiT/SiT model to better capture the overall musical structure of the song. 2) Timestamped lyric data (.LRC) is not required during the training phase; regular lyric data (.TXT) suffices.

2) Fine-tuning Phase. Finally, for the task of colloquial description to song generation, we fine-tuned the DiT/SiT model based on MuChin, as shown in **Figure 1**. In fine-tuning, the model uses type labels as conditions to process professional and amateur inputs. This fine-tuning enables the model to generate songs that meet human requirements based on user-provided colloquial descriptions and structured lyrics.

To maintain consistency between training and inference, we addressed the issue of varying input lengths, from single words to full paragraphs. During training, we randomly segmented the description annotations in MuChin into fragments of different lengths. Each fragment—whether a word, phrase, or sentence—is converted into a vector via ChinMu and concatenated along the length dimension, while keeping the hidden dimension unchanged. This method allows the DiT/SiT model to efficiently process text inputs of varying lengths during inference.

3.2 MuDiT/MuSiT Architecture

3.2.1 Architecture Overview.

We propose an end-to-end single-stage generation model named MuDiT/MuSiT, as shown in Figure 2(a). The model's primary input is colloquial descriptions; ChinMu converts these descriptions into embedding vectors, which are then concatenated with random noise to be learned via self-attention mechanisms in the DiT/SiT model. The DiT (Peebles and Xie, 2023) and SiT (Ma et al., 2024) models generate songs consistent with user descriptions, outputting VAE latent vectors. A VAE decoder decodes these latent vectors into high-quality WAV files, ensuring audio fidelity and richness.

For songs that require lyrics, structured lyrics can be generated from the same colloquial descriptions by a fine-tuned Large Language Model (LLM), or provided by the user. These lyrics, when available, can then serve as an additional conditional input to the DiT/SiT model, typically processed through its cross-attention modules.

3.2.2 Transforming Noise to VAE Space.

In original audio data, each point represents meaningful musical content. We aim to simplify this complex, high-dimensional, irregular space into a regular, normal distribution-like space, which requires a spatial transformation. We first perform unsupervised VAE pre-training on a vast dataset, converting audio into VAE latent vectors. These vectors index audio content in a VAE space that resembles a normal distribution. The VAE space is discontinuous, with noise distributed in meaningless positions between discrete points. During the DiT/SiT inference process shown in **Figure 2(b)**, we iteratively reduce noise to gradually approach a meaningful point in the VAE space, which the decoder then converts into a waveform. This means we need to make the DiT training data a subset of the VAE data.

3.2.3 Control Conditions for DiT/SiT.

We apply DiT/SiT to the song generation task, adopting this new standardized architecture to open up more possibilities for cross-domain research. **Application of Self-Attention Mechanism.** 1) Professional and amateur text descriptions. ChinMu converts text into embedding vectors. These vectors, after normalization, are concatenated with noisy samples to form noised latent vectors. These latent vectors are processed by the multi-head self-attention mechanism in DiT/SiT, thereby capturing the dependencies between text descriptions and audio content. 2) Reference audio. The noisy sample in **Figure 2(c)** is divided into prompt and target parts. The audio prompt can include user-provided reference tracks for style or content control. Source separation technology is employed to extract vocals, drums, chords, and bass as pre-training data for DiT/SiT, thereby enabling the aforementioned control. We found that even non-musical sounds can be integrated. **Application of Cross-Attention Mechanism.** 1) Lyrics and musical segment structure. The cross-attention mechanism treats variable-length lyrics and audio as parallel streams to find correlations, and maps musical segment labels as tokens to maintain temporal relationships.

3.2.4 Differential Benefits of SiT over DiT

DiT and SiT, open-sourced by Meta, use similar neural network structures. SiT enhances the interpolation and sampling processes (interpolation involves the data transformation from raw data to Gaussian noise), providing a more nuanced exploration of the diffusion process. DiT uses discrete timesteps, assuming a constant distribution conversion rate, which lacks flexibility. SiT uses continuous time, allowing for more adaptive connections between data and Gaussian noise through better interpolation functions. This aligns better with the continuity of music. Additionally, DiT uses deterministic sampling, while SiT introduces randomness, separating inference and training diffusion coefficients. This reduces overfitting and improves generalization capability. We introduce a new continuous interpolation function based on the Bezier curve. Specifically, for the original data $x^* \sim p(x)$, the transformation x_t at an arbitrary time t can be expressed as:

$$x_t = \mathbf{B}_t(x^*, \mathbf{Q}, \epsilon) = (1-t)^2 x^* + 2(1-t)t\mathbf{Q} + t^2\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0,1)$ represents random noise, and $\mathbf{Q}$ is a control point which can be defined as:

$$\mathbf{Q} = \cos\left(\frac{\pi\tilde{t}}{2}\right) x^*, \tag{2}$$

where $\tilde{t}$ is a hyperparameter. SiT uses reverse-time stochastic differential equations to relate the velocity field and the score function, allowing for the estimation of complex score functions via simpler velocity fields, thereby enhancing flexibility in music generation.

4 Experiments

4.1 Implementation Details

For the training of the ChinMu cross-modal encoder, the ChinMu vector is set to 512 dimensions. Training was conducted on a machine equipped with 80GB of VRAM and eight A800 GPUs connected via NVLink. The total effective training duration was approximately 14 days. For the pre-training of the VAE encoder and decoder, the VAE latent variable was set to 96 dimensions. We used a de-duplicated dataset of 1.5 million songs combined with MuChin as the training audio data. Training was performed on a cluster of machines equipped with A100 GPUs, totaling 64 GPUs, and took 22 days. Notably, each machine was equipped with 80GB of VRAM, featuring eight GPUs connected via NVLink, and these eight machines were interconnected via an InfiniBand (IB) network. For the pre-training of DiT or SiT, the model parameter size is 380MB, the input dimension is set to 96 dimensions, and the RoPE method is used for positional embedding. We used the same data and machine cluster as for the VAE pre-training. The total effective training duration was approximately 35 days. For the fine-tuning of DiT or SiT, the parameters and machine cluster remained consistent with

Objective		Open Source					Ours	
		StableAudio	MusicGen	Melodist	A-LDMv1	A-LDMv2	MuDiT	MuSiT
Quality (FAD)	MS-Clap↓	649.37	595.88	303.43	956.30	502.89	210.18	199.31
	Laion-Clap↓	0.85	1.16	0.78	1.32	1.07	0.55	0.48
	Encodec↓	28.60	73.84	41.34	63.62	68.22	26.43	24.83
	MERT↓	7.33	23.87	20.59	25.01	19.48	5.12	4.92
Alignment	SST↑	0.30	0.10	0.27	0.14	0.08	0.44	0.49
	ASOA↑	0.31	0.33	0.35	0.33	0.32	0.57	0.46

Table 1: Objective results. Underlined values indicate best performance.

Subjective		Open Source					Ours	
		StableAudio	MusicGen	Melodist	A-LDMv1	A-LDMv2	MuDiT	MuSiT
Quality	Vocal↑	2.16	2.23	3.98	3.23	3.37	3.96	3.95
	Accompaniment↑	3.23	3.14	3.36	3.10	3.12	4.03	4.01
	Structure↑	2.63	3.15	3.31	3.40	3.41	3.48	3.51
	Coherence↑	3.05	3.22	3.82	3.63	3.32	3.99	4.01
Alignment (Comprehension)	Lyric↑	2.74	2.61	3.25	3.08	3.43	4.26	4.35
	Genre↑	2.69	2.53	2.98	3.34	3.85	4.08	4.01
	Emotion↑	2.87	3.61	2.76	3.27	3.35	3.91	3.97
	Instrument↑	2.74	3.44	3.19	3.32	3.41	3.74	3.85

Table 2: Subjective results. Underlined values indicate best performance. (All subjective metrics exhibit statistically significant differences, $p < 0.03$.)

those used in the pre-training phase. We used the MuChin dataset, which was randomly segmented to form 300,000 text-audio paired fragments. The total effective training duration was approximately 10 days.

4.2 Main Experiment: MuDiT/MuSiT Evaluation

In this experiment, we created 50 songs using our models and baseline models. The MuDiT/MuSiT model distinguishes between professional and amateur descriptions through label control. Test data including amateur and professional descriptions, is randomly sampled from unseen data, ensuring no overlap between the ground truth song audio in the test and training sets to guarantee fairness. Each description was randomly segmented into 120-character prompts and paired with the corresponding ground truth audio.

Using these prompts, we generated songs with MuDiT, MuSiT, and several baseline models, including AudioLDM v1 (Liu et al., 2023a) and v2 (Liu et al., 2024), Stable-Audio-Open-1.0[8], Melodist (Zhiqing et al., 2024), and MusicGen (Copet et al., 2024). Notably, the Stable-Audio-Open-1.0 model performs poorly on non-English inputs, so we first translated the Chinese prompts into English.

Due to page limitations, presenting four separate tables for objective professional/objective amateur/-subjective professional/subjective amateur scores would be too verbose. This experiment averages the scores from professional descriptions and amateur descriptions to serve as the final score for each metric of the models/systems.

4.2.1 Evaluation Metrics.

Objective Evaluation Metrics.

For objective evaluation, we use Fréchet Audio Distance (FAD) to assess the sound quality of the generated audio, and Semantic Similarity (SST) and Acoustic Similarity (ASOA) to evaluate semantic alignment. FAD, adapted from the Fréchet Inception Distance (FID) in the image domain, has become a key metric for audio quality assessment, utilizing four embedding models: MS-CLAP, Laion-CLAP, Encodec, and MERT. FAD reflects the objective quality of generated audio by comparing it with a reference set. We selected 1,000 existing songs as benchmark audio to provide a comprehensive quality assessment.

1) **MS-CLAP (Elizalde et al., 2023):** Assesses overall audio quality by capturing various acoustic features. 2) **Laion-CLAP [9]:** Provides a comprehensive evaluation of audio characteristics. 3)

[8] https://www.stableaudio.com

[9] https://huggingface.co/laion/larger_clap_music

Training Data		Laion-Clap	ChinMu-Batch			ChinMu-MiniBatch		
Test Data			S-M-L	M-L	Short	S-M-L	M-L	Long
Short	Top-5↑	16.07%	14.56%	13.81%	17.87%	21.32%	21.77%	**24.02%**
	Top-10↑	26.13%	26.58%	22.07%	30.78%	37.69%	37.09%	**38.29%**
	Top-15↑	35.14%	33.78%	22.83%	39.94%	50.00%	48.20%	**50.75%**
Medium	Top-5↑	20.12%	18.92%	15.02%	17.12%	24.17%	26.58%	**27.93%**
	Top-10↑	32.88%	31.08%	27.03%	29.58%	39.79%	**41.59%**	40.39%
	Top-15↑	41.44%	43.99%	37.24%	39.64%	53.00%	50.90%	**53.15%**
Long	Top-5↑	17.12%	13.36%	13.81%	13.66%	18.92%	20.42%	**20.72%**
	Top-10↑	24.92%	24.62%	22.67%	22.52%	28.53%	**33.78%**	32.59%
	Top-15↑	31.53%	31.53%	30.48%	28.83%	38.14%	41.29%	**41.30%**
S-M-L	Top-5↑	17.77%	15.61%	14.21%	16.22%	21.47%	22.92%	**24.22%**
	Top-10↑	27.98%	27.43%	23.92%	27.63%	35.34%	**37.49%**	37.09%
	Top-15↑	36.04%	36.43%	30.18%	36.14%	47.05%	46.80%	**48.4%**

Table 3: Objective evaluation results of ChinMu with different settings. S refers to word-level annotations, such as labels and tags. M refers to short phrase descriptions. L refers to long sentence descriptions. We conduct tests on different test dataset settings to evaluate ChinMu's performance, with K values for Top-K ranking set at 5, 10, and 15. For each test setting, the best result is highlighted in bold.

Encodec (Défossez et al., 2022): Focuses on reflecting acoustic distortions and low-pass filtering effects. 4) **MERT (Li et al., 2023):** Filters out samples composed of synthetic tones with minimal temporal or spectral variation.

Furthermore, to evaluate the alignment of the generated audio, we used two methods:

1) **Semantic Similarity (SST):** This method uses the Laion-CLAP model to evaluate the alignment between the generated audio and the input text. The model maps both modalities to a shared embedding space. The Semantic Similarity score is calculated by measuring the cosine similarity between the input text vector and the corresponding audio embedding. 2) **Acoustic Similarity (ASOA):** This method evaluates the alignment between the generated audio and the ground truth audio. We use the MERT-v1-95M model (Li et al., 2023) to obtain embeddings for both the generated audio and the reference audio. The Acoustic Similarity score is derived from the cosine similarity between these embeddings.

Subjective Evaluation Metrics and Participants.

For subjective evaluation, we designed a survey questionnaire and invited 32 participants from diverse backgrounds, including 17 professional users and 15 amateur users. Songs generated by our models and baseline models were shuffled and distributed to the participants, who evaluated the quality of the music and its conformance to the input text requirements based on eight metrics. To validate these assessments, we employed t-tests for statistical data comparison. For music generation quality, we adopted four metrics:

1) **Vocal.** The sound quality of the vocals, as well as melodic and rhythmic features. 2) **Accompaniment.** Arrangement structure, instrument usage, and degree of fusion with the melody. 3) **Structure.** Musical segmentation, structural repetition, and hierarchical progression. 4) **Impression.** Impression of sound quality and coherence.

For description comprehension, we used the following metrics to measure whether the model met the input requirements:

1) **Lyrics.** Alignment of lyrics with the input. 2) **Genre.** Consistency of style with the input. 3) **Emotion.** Emotional resonance with the input. 4) **Instrument.** Consistency of instruments with the specified input.

4.2.2 Results

Objective Evaluation Results.

MuSiT and MuDiT were compared with baseline music generation models. As shown in Table 1, MuSiT and MuDiT outperform these baselines on key metrics such as FAD. Furthermore, MuDiT/MuSiT achieved the highest alignment scores. While Tables 1 and 2 average scores for brevity, MuDiT/MuSiT demonstrated strong individual performance on both professional and amateur descriptions when evaluated separately, confirming its efficacy across diverse user types.

Notably, MuSiT excels in balancing generation quality and alignment, adeptly handling descriptive music generation tasks. These findings support the hypothesis that the SiT architecture, tailored for temporal data, is more effective for music generation—especially for continuous and dynamic elements—than the DiT-based MuDiT. Although MuDiT also aligns well with user prompts, MuSiT emerges as the superior choice for tasks demanding high fidelity and precise alignment with colloquial musical descriptions.

Subjective Evaluation Results.

The subjective evaluation results in **Table 2** corroborate the objective findings. In music generation quality, MuSiT and MuDiT slightly outperform other baseline models. For understanding and aligning with professional and amateur colloquial descriptions, MuDiT and MuDiT far surpass other baseline models. Although MuDiT's performance in understanding colloquial descriptions is comparable to MuSiT's, given MuSiT's advantage in music generation quality, "our model" specifically refers to MuSiT in subsequent experiments.

4.3 Method Analysis: ChinMu Cross-Modal Encoder

By comparing the performance differences of ChinMu with training data of different length levels and under different training strategies, we aim to explore the optimal performance of ChinMu. We divide the text descriptions in MuChin into three annotation length levels: **S**hort annotations, including word-level labels and tags; **M**edium annotations, consisting of short phrase descriptions; and **L**ong annotations, providing more detailed long sentence descriptions. In addition to using descriptions of different lengths to train our model, we also experimented with training strategies including MiniBatch and Batch. Batch training updates parameters using the entire dataset, while MiniBatch training uses subsets of the dataset for each update.

4.3.1 Evaluation Metrics.

We designed a retrieval-based metric similar to (Xu et al., 2018). Given a description, we calculate the top-K retrieval accuracy from a pool of N candidate audios. The process begins by constructing a pool of N candidate audios, obtained by applying the ChinMu model to the audio data in the test set. Subsequently, we compute and rank the cosine similarity between the given description and each of the N candidate audio samples. If the ground truth audio GT is ranked in the top-K positions, we consider the retrieval successful. Given M distinct descriptions, the retrieval accuracy RA is defined as:

$$f_n = \begin{cases} 1, & GT \text{ in top-K positions} \\ 0, & GT \text{ not in top-K positions} \end{cases} \tag{3}$$

$$RA = \frac{1}{M} \cdot \sum_{i=1}^{M} f_n(i) \tag{4}$$

where $f_n(i)$ indicates whether the i-th description was successfully retrieved.

4.3.2 Results.

ChinMu's performance was evaluated across different configurations and test datasets using Top-K retrieval accuracy (K values of 5, 10, and 15). **Table 3** shows that ChinMu outperforms existing pre-trained models (like Laion-CLAP) on unseen colloquial description-audio test sets. This suggests ChinMu has a stronger understanding of colloquial descriptions, potentially enabling MuDiT and MuSiT to better comprehend user requests.

ChinMu-MiniBatch consistently outperformed ChinMu-Batch across various evaluation metrics. This can be attributed to more frequent parameter updates in MiniBatch training, allowing for a more nuanced adaptation to the dynamic nature of colloquial descriptions. MiniBatch training also introduces more randomness and accelerates convergence, leading to better generalization across different input styles.

Finally, ChinMu demonstrates superior performance on long descriptions compared to medium and short ones, and on the S-M-L dataset mixing all three types. This is likely due to the richer semantic information in long descriptions providing more contextual clues for better alignment. The additional detail enables ChinMu to generate outputs more accurately aligned with the given descriptions, whereas shorter inputs may lack necessary context for precise mapping.

5 Conclusion

Our work represents a significant advancement in the field of human-AI collaborative music generation. The MuDiT/MuSiT framework has demonstrated the potential to combine AI-generated music with professional and amateur colloquial expressions, paving the way for more inclusive and nuanced AI-generated artistic creations. Future work will explore expanding datasets and applying our methods to other creative domains.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62272409).

Limitations

While our MuDiT/MuSiT framework represents a notable advancement in aligning generative music models with user expectations, particularly for Chinese colloquial descriptions, we acknowledge several limitations that warrant consideration for future research.

First, our framework, MuDiT/MuSiT, is currently tailored to Chinese language and musical contexts. Key components like our ChinMu cross-modal encoder (built upon architectures such as CLAP (Wu et al., 2023b) and MuLan (Huang et al., 2022)) and our fine-tuned lyric LLM (Qwen-14B (Bai et al., 2023)) were trained primarily using the MuChin dataset (Wang et al., 2024b). While effective for this domain, extending MuDiT/MuSiT to other languages and musical cultures would require significant new datasets and retraining efforts.

Second, our lyric generation and integration process offers room for improvement. Our use of Automatic Speech Recognition (ASR) for some training data lyrics may introduce errors that could affect model learning. Additionally, while our system incorporates structured lyrics, for instance, from our fine-tuned LLM (Qwen-14B (Bai et al., 2023)), ensuring seamless and expressive synchronization with complex musical elements remains a challenge.

Third, our decision to generate entire songs at once, rather than using window-based methods, aids in capturing overall musical structure. However, this approach may face scalability and coherence issues for very long songs. The impact of padding, used by us to standardize audio lengths, on such pieces—especially regarding computational efficiency and sustained musical interest—requires further study.

Finally, while MuDiT/MuSiT demonstrates enhanced alignment with user descriptions and supports reference audio, improving nuanced musical controllability is an avenue for future work. For example, playing a specific instrument at a particular position, responding to harmonic developments, or achieving deeply nuanced emotional delivery in vocals—to allow for more sophisticated and detailed user guidance, remains a key challenge.

References

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

Tarek Amer, Beste Kalender, Lynn Hasher, Sandra E Trehub, and Yukwal Wong. 2013. Do older professional musicians have cognitive advantages? *PloS one*, 8(8):e71630.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ye Bai, Haonan Chen, Jitong Chen, Zhuo Chen, Yi Deng, Xiaohong Dong, Lamtharn Hantrakul, Weituo Hao, Qingqing Huang, Zhongyi Huang, Dongya Jia, Feihu La, Duc Le, Bochen Li, Chumin Li, Hui Li, Xingxing Li, Shouda Liu, Wei-Tsung Lu, Yiqing Lu, Andrew Shaw, Janne Spijkervet, Yakun Sun, Bo Wang, Ju-Chiang Wang, Yuping Wang, Yuxuan Wang, Ling Xu, Yifeng Yang, Chao Yao, Shuo Zhang, Yang Zhang, Yilin Zhang, Hang Zhao, Ziyi Zhao, Dejian Zhong, Shicen Zhou, and Pei Zou. 2024. Seed-music: A unified framework for high quality and controlled music generation. *Preprint*, arXiv:2409.09214.

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

Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. ICML.

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

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

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

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2024. Stable audio open. *arXiv preprint arXiv:2407.14358*.

Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. 2021. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 178–186.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*.

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

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Shun Lei, Yixuan Zhou, Boshi Tang, Max W. Y. Lam, Feng Liu, Hangyu Liu, Jingcheng Wu, Shiyin Kang, Zhiyong Wu, and Helen Meng. 2024. Songcreator: Lyrics-based universal song generation. *Preprint*, arXiv:2409.06029.

Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R Manmatha, Ashwin Swaminathan, Zhuowen Tu, Stefano Ermon, and Stefano Soatto. 2024. On the scalability of diffusion-based text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9400–9409.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Yike Guo, and Jie Fu. 2023. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*, arXiv:2306.00107.

Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. 2025. Musiceval: A generative music corpus with expert ratings for automatic text-to-music evaluation. *Preprint*, arXiv:2501.10811.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

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

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

Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. 2023. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. 2024. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*.

Weihang Mao, Bo Han, and Zihao Wang. 2023. Sketchffusion: Sketch-guided image editing with diffusion model. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 790–794.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2023. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*.

Christian Alexander Mikutta, Gieri Maissen, Andreas Altorfer, Werner Strik, and Thomas Koenig. 2014. Professional musicians listen differently to music. *Neuroscience*, 268:102–111.

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

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

Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13798–13805.

Zihao Wang, Ma Le, Liu Yan, and Zhang Kejun. 2024a. Samoye: Zero-shot singing voice conversion based on feature disentanglement and synthesis. *arXiv preprint arXiv:2407.07728*.

Zihao Wang, Shuyu Li, Tao Zhang, Qi Wang, Pengfei Yu, Jinyang Luo, Yan Liu, Ming Xi, and Kejun Zhang. 2024b. Muchin: A chinese colloquial description benchmark for evaluating language models in the field of music. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7771–7779.

Zihao Wang, Le Ma, Chen Zhang, Bo Han, Yunfei Xu, Yikai Wang, Xinyi Chen, Haorong Hong, Wenbo Liu, Xinda Wu, and Kejun Zhang. 2024c. Remast: Real-time emotion-based music arrangement with soft transition. *IEEE Transactions on Affective Computing*, pages 1–15.

Zihao Wang, Kejun Zhang, Yuxing Wang, Chen Zhang, Qihao Liang, Pengfei Yu, Yongsheng Feng, Wenbo Liu, Yikai Wang, Yuntao Bao, et al. 2022. Songdriver: Real-time music accompaniment generation without logical latency nor exposure bias. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1057–1067.

Xinda Wu, Zhijie Huang, Kejun Zhang, Jiaxing Yu, Xu Tan, Tieyao Zhang, Zihao Wang, and Lingyun Sun. 2023a. Melodyglm: multi-task pre-training for symbolic melody generation. *arXiv preprint arXiv:2309.10738*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023b. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *Preprint*, arXiv:2402.07729.

Botao Yu, Peiling Lu, Rui Wang, Wei Hu, Xu Tan, Wei Ye, Shikun Zhang, Tao Qin, and Tie-Yan Liu. 2022. Museformer: Transformer with fine-and coarse-grained attention for music generation. *Advances in Neural Information Processing Systems*, 35:1376–1388.

Jiaxing Yu, Songruoyao Wu, Guanting Lu, Zijin Li, Li Zhou, and Kejun Zhang. 2024a. Suno: potential, prospects, and trends. *Frontiers of Information Technology & Electronic Engineering*, pages 1–6.

Jiaxing Yu, Xinda Wu, Yunfei Xu, Tieyao Zhang, Songruoyao Wu, Le Ma, and Kejun Zhang. 2024b. Songglm: Lyric-to-melody generation with 2d alignment encoding and multi-task pre-training. *arXiv preprint arXiv:2412.18107*.

Ruibin Yuan, Hanfeng Lin, Shawn Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Xingjian Du, Xeron Du, Zhen Ye, Tianyu Zheng, Yinghao Ma, Minghao Liu, Lijun Yu, Zeyue Tian, Ziya Zhou, Liumeng Xue, Xingwei Qu, Yizhi Li, Tianhao Shen, Ziyang Ma, Shangda Wu, Jun Zhan, Chunhui Wang, Yatian Wang, Xiaohuan Zhou, Xiaowei Chi, Xinyue Zhang, Zhenzhu Yang, Yiming Liang, Xiangzhou Wang, Shansong Liu, Lingrui Mei, Peng Li, Yong Chen, Chenghua Lin, Xie Chen, Gus Xia, Zhaoxiang Zhang, Chao Zhang, Wenhu Chen, Xinyu Zhou, Xipeng Qiu, Roger Dannenberg, Jiaheng Liu, Jian Yang, Stephen Huang, Wei Xue, Xu Tan, and Yike Guo. 2025. Yue: Open music foundation models for full-song generation. `https://github.com/multimodal-art-projection/YuE`. GitHub repository.

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

Hong Zhiqing, Huang Rongjie, Cheng Xize, Wang Yongqi, Li Ruiqi, You Fuming, Zhao Zhou, and Zhang Zhimeng. 2024. Text-to-song: Towards controllable music generation incorporating vocals and accompaniment. *arXiv preprint arXiv:2404.09313*.

Pengfei Zhu, Chao Pang, Yekun Chai, Lei Li, Shuohuan Wang, Yu Sun, Hao Tian, and Hua Wu. 2023. Erniemusic: Text-to-waveform music generation with diffusion models. *arXiv preprint arXiv:2302.04456*.