ERNIE-Music: Text-to-Waveform Music Generation with Diffusion Models

Pengfei Zhu, Chao Pang, Yekun Chai, Lei Li, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu
Baidu Inc.
{zhupengfei03, pangchao04, chaiyekun, wangshuohuan, sunyu02}@baidu.com

Abstract

In recent years, the burgeoning interest in diffusion models has led to significant advances in image and speech generation. Nevertheless, the direct synthesis of music waveforms from unrestricted textual prompts remains a relatively underexplored domain. In response to this lacuna, this paper introduces a pioneering contribution in the form of a text-to-waveform music generation model, underpinned by the utilization of diffusion models. Our methodology hinges on the innovative incorporation of free-form textual prompts as conditional factors to guide the waveform generation process within the diffusion model framework. Addressing the challenge of limited text-music parallel data, we undertake the creation of a dataset by harnessing web resources, a task facilitated by weak supervision techniques. Furthermore, a rigorous empirical inquiry is undertaken to contrast the efficacy of two distinct prompt formats for text conditioning, namely, music tags and unconstrained textual descriptions. The outcomes of this comparative analysis affirm the superior performance of our proposed model in terms of enhancing text-music relevance. Finally, our work culminates in a demonstrative exhibition of the excellent capabilities of our model in text-to-music generation. We further demonstrate that our generated music in the waveform domain outperforms previous works by a large margin in terms of diversity, quality, and text-music relevance.[1]

1 Introduction

Music, as a sophisticated and profound human art form, possesses a unique capacity to evoke emotions, alter moods, and tell compelling stories through its intricate interplay of harmony, melody, and rhythm. In recent years, the realm of music generation has garnered significant attention and interest, coinciding with the rapid advancements in deep learning techniques.

Within this context, some research endeavors, exemplified by works such as (Wu and Sun, 2022), have concentrated on the domain of symbolic music generation. This approach entails the acquisition of knowledge to predict sequences of musical composition, encompassing elements such as notes, pitch, and dynamic attributes. However, it is noteworthy that the resultant symbolic music lacks performance attributes, necessitating subsequent post-processing to synthesize the auditory experience of the musical piece. Conversely, an alternative line of inquiry, exemplified by works such as (Pasini and Schlüter, 2022), has been dedicated to the generation of audio or waveform-based music. Notably, this approach obviates the need for additional synthesis steps, as it directly produces audio signals. Nevertheless, it is important to recognize that generating audio signals in this manner often presents inherent challenges in controlling and fine-tuning performance attributes to achieve the desired level of quality and satisfaction.

Besides works on unconditional music generation, there have been explorations about conditional music generation (Pasini and Schlüter, 2022; Zhuo et al., 2022), which aims to meet the application requirements in scenarios such as automatic video soundtrack creation and music creation with specific genres or features. Notably, generative models can leverage information from various modalities, such as text and image, to create relevant outputs for a conditional generation.

In addition to unconditional music generation, there is a growing interest in conditional music generation (Pasini and Schlüter, 2022; Zhuo et al., 2022). This field caters to specific application needs, like creating video soundtracks or generating music with specific genres or features. Generative models can use various data modalities, such as text and images, to create relevant outputs in con-

[1]Generated cases are available at https://reurl.cc/94W4yO

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations, pages 86–95
November 1–4, 2023.

ditional music generation. Nonetheless, the challenge of directly generating musical waveforms from unrestricted textual input remains a relatively underexplored frontier. While research efforts have delved into text-conditioned music generation, exemplified by works such as (Wu and Sun, 2022), BUTTER (Zhang et al., 2020), and Mubert[2], it is noteworthy that these approaches do not possess the capability to directly produce musical audio based on unstructured free-form text prompts.

To address prior limitations, we introduce ERNIE-Music, a pioneering effort in free-form text-to-music generation using diffusion models in the waveform domain. To overcome the shortage of parallel text-to-music data, we have undertaken the collection of music waveforms along with their corresponding top-voted comments from the internet. We employ conditional diffusion models to generate musical waveforms and investigate the impact of text format on enhancing text-music relevance.

To conclude, the contributions of this paper are:

- We introduce a music generation model that leverages free-form text as a conditioning factor, utilizing the diffusion model to generate waveform-based music.
- We curate a dataset of free-form text-music parallel data from the internet.
- We investigate and compare the efficacy of two text formats for conditioning the generative model, demonstrating that the use of free-form text significantly enhances text-music relevance.
- Our results highlight the model's ability to produce diverse, high-quality music with markedly improved text-music relevance, surpassing existing methods by a large margin.

2 Related Work

Controllable Music Generation Controlled music generation faces the persistent challenge of effectively imposing constraints on musical output. Previous approaches have employed various techniques to address this issue. For instance, VQ-CPC (Hadjeres and Crestel, 2020) focuses on learning local music features devoid of temporal information. Meanwhile, (Pasini and Schlüter, 2022) leverages tempo information as a condition for generating music in the "techno" genre. BUTTER (Zhang et al., 2020) adopts a natural language representation encompassing attributes like music key, meter, and style to exercise control over music generation. Furthermore, (Wu and Sun, 2022) extends this exploration by investigating the impact of different pre-trained models in text-to-music generation. Besides, retrieval-based methods can be adopted to generate music by combining human-created music pieces. Mubert firstly employs Sentence-BERT (Reimers and Gurevych, 2019) to encode input natural language and music tags, secondly retrieves relevant tags based on the distance among the representation, finally combines relevant sounds (all crafted by musicians and sound designers) to obtain the generated music.

Diffusion Models Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) are latent variable models rooted in non-equilibrium thermodynamics. They operate by gradually disassembling the structure of the original data distribution through a progressive forward diffusion process and subsequently acquire the means to reconstruct the original data via a finite iterative denoising process. In recent years, diffusion models have gained significant traction across diverse domains, including image generation (Nichol et al., 2022; Dhariwal and Nichol, 2021; Ramesh et al., 2022) and audio generation (Chen et al., 2021; Kreuk et al., 2022). Our work is closely aligned with the realm of diffusion-based approaches in text-to-audio generation (Chen et al., 2021; Kreuk et al., 2022). Some concurrent works (Huang et al., 2023; Schneider et al., 2023; Agostinelli et al., 2023) use diffusion models to tackle the text-to-music generation, which mainly focus on given specific music genres or instruments. It is important to note that while these prior works primarily focus on speech generation or some restricted text descriptions, our research extends the application of diffusion models to the synthesis of music waveforms based on arbitrary textual prompts, representing a distinct task within the audio generation domain.

3 Method

This section commences with an overview of diffusion models, providing the overall context for our subsequent discussion. Subsequently, we delve into the specifics of our text-conditional diffusion model, elucidating its architecture and the objectives underpinning its training.

[2]https://github.com/MubertAI/Mubert-Text-to-Music

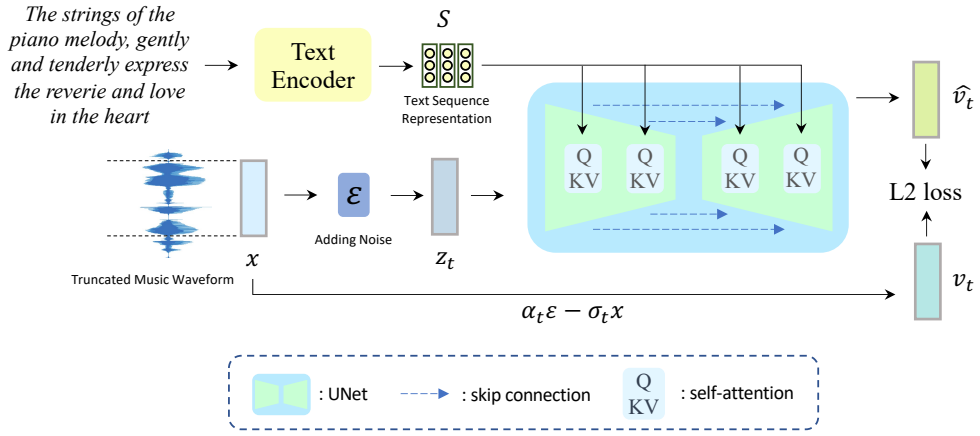

Figure 1: The overall architecture of text-to-music generation training. The text is input to the text encoder to obtain the sequence representation S, then S and the sampled music waveform (noise added) z_t are input to the UNet to obtain the estimated volocity $\hat{v}_t$, finally we calculate the L2 loss between $\hat{v}_t$ and the real volocity v_t. For the input text, the original Chinese is "钢琴旋律的弦音，轻轻地、温柔地倾诉心中的遐想、心中的爱恋".

3.1 Unconditional Diffusion Model

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) are composed of two essential components: a *forward process*, where noise is progressively incorporated into a data sample, and a *reverse process*, which subsequently removes this noise through multiple iterations to produce a sample that aligns with the authentic data distribution. Specifically, our approach is founded upon the diffusion model formulated within a continuous-time framework (Kingma et al., 2021; Tzen and Raginsky, 2019; Chen et al., 2021; Song et al., 2021; Salimans and Ho, 2022).

In the context of diffusion models, we begin with a data sample denoted as x drawn from the distribution $p(x)$. These models make use of latent variables z_t, where the parameter t spans the continuous interval from 0 to 1. The log signal-to-noise ratio, represented as λ_t, is precisely defined as $\lambda_t = \log\left(\frac{\alpha_t^2}{\sigma_t^2}\right)$, where α_t and σ_t correspond to the components of the noise schedule.

During the *forward process*, often referred to as the *diffusion process*, we progressively incorporate Gaussian noise into the sample, conforming to a Markov chain, characterized by the following progression:

$$q(z_t|x) = \mathcal{N}(z_t; \alpha_t x, \sigma_t^2 \mathbf{I}) \quad (1)$$

$$q(z_{t'}|z_t) = \mathcal{N}(z_{t'}; (\alpha_{t'}/\alpha_t)z_t, \sigma_{t'|t}^2 \mathbf{I}) \quad (2)$$

where $t, t' \in [0,1]$ and $t < t'$, and $\sigma_{t'|t}^2 = (1 - e^{\lambda_{t'} - \lambda_t})\sigma_{t'}^2$.

In the *reverse process*, a function approximation with parameters θ (denoted as $\hat{x}_\theta(z_t, \lambda_t, t) \approx x$) estimates the denoising procedure:

$$p_\theta(z_t|z_{t'}) = \mathcal{N}(z_t; \tilde{\mu}_{t|t'}(z_{t'}, x)), \tilde{\sigma}_{t|t'}^2 \mathbf{I}) \quad (3)$$

where $\tilde{\mu}_{t|t'}(z_{t'}, x, t')) = e^{\lambda_{t'} - \lambda_t}(\alpha_t/\alpha_{t'})z_{t'} + (1 - e^{\lambda_{t'} - \lambda_t})\alpha_t x$.

Initiating from $z_1 \sim \mathcal{N}(0, \mathbf{I})$, the *reverse process* involves the sequential application of the denoising procedure to the latent variables z_t, ultimately yielding $z_0 = \hat{x}$. To train the denoising model $\hat{x}_\theta(z_t, \lambda_t, t)$, we adopt the weighted mean squared error loss as our optimization objective:

$$L = E_{t\sim[0,1],\epsilon\sim\mathcal{N}(0,\mathbf{I})}[w(\lambda_t)\|\hat{x}_\theta(z_t, \lambda_t, t) - x\|_2^2] \quad (4)$$

where $w(\lambda_t)$ denotes the weighting function and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ denotes the noise.

3.2 Conditional Diffusion Model

Numerous studies have effectively employed generative models within conditional settings (Mirza and Osindero, 2014; Sohn et al., 2015; Rombach et al., 2022). In the case of conditional diffusion models, the focus shifts from approximating the distribution $p(x)$ to $p(x|y)$, achieved by modeling the denoising process as $\hat{x}_\theta(z_t, \lambda_t, t, y)$, where y represents the conditioning variable. This conditioning variable y can assume various modalities, encompassing image, text, and audio.

In the text-to-music generation scenario, y takes the form of a textual prompt. This textual input

serves as a guiding element for the model, steering it toward the generation of music that corresponds to the provided text. In subsequent sections, we delve into the intricacies of modeling the conditional diffusion model in detail.

3.2.1 Model Architecture

For text-to-music generation, our diffusion process conditions on textual input denoted as y. As illustrated in Figure 1, our comprehensive model architecture comprises a conditional music diffusion model responsible for modeling the anticipated *velocity*, represented as $\hat{v}_\theta(z_t, t, y)$ (Salimans and Ho, 2022). Additionally, we incorporate a text encoder denoted as $\mathrm{E}(\cdot)$, which transforms text tokens with a length of n into a sequence of vector representations $[s_0; S]$, each possessing a dimensionality of d_E. Here, $S = [s_1, ..., s_n]$, with $s_i \in \mathbb{R}^{d_E}$, and s_0 serving as the classification representation of the input text.

The inputs to the music diffusion model encompass the latent variable $z_t \in \mathbb{R}^{d_c \times d_s}$, the timestep t (which is subsequently transformed into the embedding $e_t \in \mathbb{R}^{d_t \times d_s}$), and the representation of the text sequence $[s_0; S] \in \mathbb{R}^{(n+1)\times d_E}$. Here, d_c corresponds to the number of channels, d_s signifies the sample size, and d_t denotes the feature size of the timestep embedding. The output of this architecture is represented by the estimated *velocity*, denoted as $\hat{v}_t \in \mathbb{R}^{d_c \times d_s}$.

Inspired by previous works on latent diffusion models (Nichol et al., 2022; Rombach et al., 2022; Dhariwal and Nichol, 2021), we have adopted the architecture of UNet (Ronneberger et al., 2015) whose key components are stacked convolutional blocks and self-attention blocks (Vaswani et al., 2017). Generation models can estimate the conditional distribution, notably $p(x|y)$, and there exist various techniques to integrate conditional information y into generative models (Sohn et al., 2015).

Our diffusion network is designed to predict the latent velocity, denoted as $\hat{v}_\theta$, at randomly sampled timestep t, leveraging the noised latent z_t and a textual input $[s_0; S]$ as conditioning elements. To integrate the conditioning information into the diffusion process, we employ a fusion operation, denoted as $\mathrm{Fuse}(\cdot, \cdot)$, on the timestep embedding e_t and the text classification representation s_0. This operation yields a text-aware timestep embedding, $e't = \mathrm{Fuse}(e_t, s_0) \in \mathbb{R}^{dt' \times d_s}$. Subsequently, we concatenate this modified embedding with z_t to derive $z'_t = (z_t \oplus e't) \in \mathbb{R}^{(dt'+d_c)\times d_s}$. It is worth noting that, for simplicity, the operations involving the timestep embedding have been omitted from Figure 1. In Section 4.6, we delve into a comparative analysis of different implementations of the fusion operation, $\mathrm{Fuse}(\cdot, \cdot)$, to evaluate their performance.

Furthermore, we incorporate the conditional representation into the self-attention blocks (Vaswani et al., 2017), which are responsible for capturing global information within the music signals. Within the self-attention blocks, taking into account the intermediate representation, where $z'_t \in \mathbb{R}^{(d_t+d_c)\times d_s}$ is denoted as $\phi(z'_t) \in \mathbb{R}^{d_a \times d_\phi}$, and $S \in \mathbb{R}^{n \times d_E}$, the output is calculated in the following manner:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (5)$$

$$\mathrm{head_i} = \mathrm{Attention}(Q_i, K_i, V_i) \quad (6)$$

$$Q_i = \phi(z'_t) \cdot W_i^Q \quad (7)$$

$$K_i = \mathrm{Concat}(\phi(z'_t) \cdot W_i^K,\ S \cdot W_i^{SK}) \quad (8)$$

$$V_i = \mathrm{Concat}(\phi(z'_t) \cdot W_i^V,\ S \cdot W_i^{SV}) \quad (9)$$

$$\mathrm{CSA}(\phi(z'_t),\ S) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head_h})W^O \quad (10)$$

where $W_i^Q \in \mathbb{R}^{d_\phi \times d_q}$, $W_i^K \in \mathbb{R}^{d_\phi \times d_k}$, $W_i^V \in \mathbb{R}^{d_\phi \times d_v}$, $W_i^{SK} \in \mathbb{R}^{d_E \times d_k}$, $W_i^{SV} \in \mathbb{R}^{d_E \times d_v}$, $W^O \in \mathbb{R}^{hd_v \times d_\phi}$ are parameter matrices, and h denotes the number of heads, and $\mathrm{CSA}(\cdot, \cdot)$ denotes the conditional self-attention operation.

3.2.2 Training

Following (Salimans and Ho, 2022), we set the weighting function in equation 4 as the "SNR+1" weighting for a more stable denoising process.

Specifically, for the noise schedule α_t and σ_t, we adopt the cosine schedule (Nichol and Dhariwal, 2021) $\alpha_t = \cos(\pi t/2)$, $\sigma_t = \sin(\pi t/2)$, and the variance-preserving diffusion process satisfies $\alpha_t^2 + \sigma_t^2 = 1$. We denote the function approximation as $\hat{v}_\theta(z_t, t, y)$, where y denotes the condition. The prediction target of $\hat{v}_\theta(z_t, t, y)$ is *velocity* $v_t \equiv \alpha_t \epsilon - \sigma_t x$, which gives $\hat{x} = \alpha_t z_t - \sigma_t \hat{v}_\theta(z_t, t, y)$. Finally, our training objective is:

$$L_\theta = (1 + \alpha_t^2/\sigma_t^2)\|x - \hat{x}_t\|_2^2 \quad (11)$$

$$= \|v_t - \hat{v}_t\|_2^2 \quad (12)$$

Algorithm 1 (in Appendix) displays the complete training process with the diffusion objective proposed by (Salimans and Ho, 2022).

	Train	Test
Num. of Data Samples	3890	204
Avg. Text (Tokens) Length	63.23	64.45
Music Sample Rate	16000	
Music Sample Size	327680	
Music Duration	20 seconds	

Table 1: The statistics of our collected dataset.

4 Experiments

4.1 Implementation Details

Following previous works (Rombach et al., 2022; Nichol et al., 2022; Ho et al., 2020), we use UNet (Ronneberger et al., 2015) architecture for the diffusion model. The UNet model uses 14 layers of stacked convolutional blocks and attention blocks for the downsample and upsample module, with skipped connections between layers with the same hidden size. It uses the input/output channels of 512 for the first ten layers and two 256s and 128s afterward. We employ the attention at 16x16, 8x8, and 4x4 resolutions. The sample size and sample rate of the waveform are 327,680 and 16,000, and the channel size is 2. The timestep embedding layer contains trainable parameters of 8x1 shape. It concatenates the noise schedule to obtain the embedding, which is then expanded to the sample size to obtain $e_t \in \mathbb{R}^{16\times 327,680}$. For the text encoder $\mathrm{E}(\cdot)$, we use ERNIE-M (Ouyang et al., 2021) to encode multi-lingual text inputs such as Chinese, English, Korean and Japanese, etc.

4.2 Dataset

Users on music platforms comments on music they like and they upvotes the comments they favor. We observe that popular comments with high upvotes is of high quality and contain useful music-related information such as musical instruments, genres, and human moods. Thus we collect a large set of text-music pairs data as training set.

The statistics of our collected Web Music with Text dataset and examples are listed in Table 1 and 7. Note that the time duration of our collected music samples are usually 2 to 3 minutes, thus the actual number of training samples may be considered as 6 to 9 times larger than the music samples because of randomly sampling 20 seconds during the training process.

4.3 Evaluation Metric

For the text-to-music generation task, we evaluate performance in two aspects: text-music relevance and music quality. Because there is currently a lack of well-recognized and authoritative objective evaluation methods for text-music relevance, and the objective metrics for evaluation music quality such as Frechet Audio Distance (FAD) only calculate the closeness between the generated music and the real music instead of the actual quality (Zhuo et al., 2022), we employ human evaluation methods. We use the compared methods or models to generate music based on texts from the test set, and manually score the generated music and calculate the mean score by averaging over results from different evaluators. We hire 10 people (who are of average listener annotator level among human) to perform human evaluation, scoring the music generated by each compared model, and then average the scores over the 10 people for each generated music. The identification of models corresponding to the generated music is invisible to the evaluators. Finally, we average the scores of the same model on the entire test samples to obtain the final evaluation results of the models.

4.4 Compared Methods

The methods for comparison are Text-to-Symbolic Music (denoted as TSM) (Wu and Sun, 2022), Mubert and Musika (Pasini and Schlüter, 2022). The generated music from Mubert is actually created by human musicians, and TSM only generates music score, which needs to be synthesized into music audio by additional tools, so the music quality among Mubert, TSM, and our model is not comparable. Thus, we only compare the text-music relevance between them and our model. To synthesize the music audio based on the symbolic music score generated by TSM, we first adopt abcMIDI[3] to convert the abc file output by TSM to MIDI file and then use FluidSynth[4] to synthesize the final music audio. For music quality, we compare our model's performance with Musika, a recent famous work that also directly generates waveform music.

4.5 Results

Table 2 and 3 show the evaluation results of text-music relevance and music quality. For text-music relevance evaluation, we use a ranking score of 3

[3]https://github.com/sshlien/abcmidi
[4]https://github.com/FluidSynth/fluidsynth

Method	Score↑	Top Rate ↑	Bottom Rate↓
TSM (Wu and Sun, 2022)	2.05	12%	27%
Mubert	1.85	37%	32%
our model	**2.43**	**55%**	**12%**

Table 2: Comparison of text-music relevance.

Method	Score↑	Top Rate↑	Bottom Rate↓
Musika (Pasini and Schlüter, 2022)	3.03	5%	13%
our model	**3.63**	**15%**	**2%**

Table 3: Comparison of music quality.

(best), 2, 1 to denote which of the three models has the best relevance given a piece of text. For music quality, we use a five-level score of 5 (best), 4, 3, 2, 1, which indicates to what extent the evaluator prefers the melody and coherence of the music. The top rate means the probability that the music obtains the highest score, and the bottom rate means the probability that the music obtains the lowest score. The results indicate that our model can generate music with better quality and text-music relevance which outperforms related works by a large margin.

4.6 Analysis

Diversity The music generated by our model has a high level of diversity. For melody, our model can generate music with a softer and more soothing rhythm or more passionate and fast-paced music. For emotional expression, some music sound sad, while some are very festive and cheerful. For musical instruments, it can generate music composed by various instruments, including piano, violin, erhu, and guitar. We select two examples with apparent differences and analyze them based on the visualization results. As shown in the waveform from Figure 2, the fast-paced guitar piece has denser sound waves, while the piano pieces have a slower, more soothing rhythm. Moreover, the spectrogram shows that the guitar piece holds dense high and low-frequency sounds, while the piano piece is mainly in the bass part.

Comparison of Different Text Condition Fusing Operations As introduced in Section 3.2.1, we compare two implementations of the fusing operation $\mathrm{Fuse}(\cdot,\cdot)$, namely concatenation and element-wise summation. To evaluate the effect, we compare the performance on the test set as the training progresses. For every 5 training steps, we adopt the model checkpoint to generate pieces of music based on the texts in the test set and calculate the Mean Squared Error (MSE) of generated music and gold music from the test set. The visualization results shown in Figure 3 indicate no apparent difference between the two fusing operations, thus we adopt the element-wise summation for simplicity.

Comparison of Different Formats of Input Text Our proposed method leverages free-form text to generate music. However, considering that the more widely used methods in other works generate music based on a set of pre-defined music tags representing the specific music's feature (Zhang et al., 2020), we compare these two methods to obtain better text-music relevance of generated music: (1) *End-to-End Text Conditioning*. Suppose the training data consists of multiple text and music pairs $< Y, X >$. The texts in Y are free-form, describing some scenario, emotion, or just a few words about music features. We adopt the straightforward way to process the texts: to input them into the text encoder $\mathrm{E}(\cdot)$ to obtain the text representations. It relies on the natural high correlation of the $< Y, X >$, and the conditional diffusion model dynamically learns to capture the critical information from the text in the training process. (2) *Music Tag Conditioning*. Using short and precise music tags as the text condition may make it easier for the model to learn the mapping between text and corresponding music. We analyze the text data from the training set and distill critical information from the texts to obtain music tags. Examples as shown in Table 5. The key features of the music in a piece of long text are limited and can be extracted as music tags. We randomly select 50 samples from the test set for manual evaluation. Table 4 shows the evaluation results of the two conditioning methods, which indicates that our proposed free-form text-based music generation method obtains better text-music relevance than using pre-defined music tags. The main reason might be that the human-made music tag selection rules introduce much noise and result in the loss of some useful information from the original text. Thus it is better to use the *End-to-End Text Conditioning* method for the model to learn to capture useful information dynamically.

5 Conclusion

In this paper, we present ERNIE-Music, a novel music generation model that directly creates music

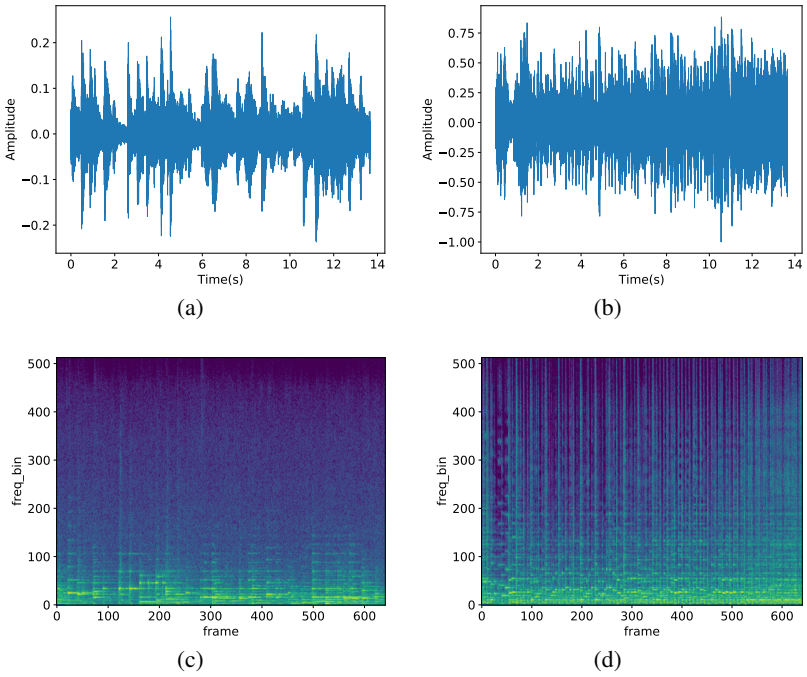

Figure 2: The spectrogram and waveform of generated music examples. The model can generate diverse music, including smoothing and cadenced (a, c) and fast-paced (b, d) rhythms. Text of (a, c): The piano piece is light and comfortable yet deeply affectionate. Text of (b, d): A passionate, fast-paced guitar piece.

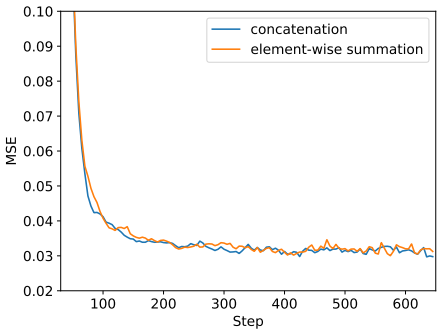

Figure 3: The MSE results on the test set for two implementations of the fusing operation.

Method	Score↑	Top Rate↑	Bottom Rate↓
Music Tag Conditioning	1.7	22%	52%
End-to-End Text Conditioning	**2.3**	**40%**	**10%**

Table 4: Comparison of text-music relevance between two conditioning text formats.

from free-form text. To overcome the scarcity of text-music parallel data, we collect music paired with descriptive comment texts from the internet. We investigate the impact of text format on text-music relevance by comparing two text conditioning methods. Our results showcase ERNIE-Music's ability to generate diverse, coherent music, outperforming existing approaches in music quality and text-music relevance.

Limitations

While our model successfully generates coherent and pleasant music, it is important to acknowledge several limitations that can be addressed in future research. The primary limitation is the fixed and relatively short length of the generated music. Due to computational resource constraints, we were unable to train the model on longer sequences. Altering the length during the inference phase can negatively impact performance, which is an area for further investigation.

Another limitation is the relatively slow speed of the generation process. The iterative nature of the generation procedure contributes to this slower speed. Exploring techniques to optimize the generation process and reduce computational overhead could enhance the efficiency of music generation in the future.

Additionally, our current model is designed to generate instrumental music and does not incorporate human voice. This limitation stems from the training data used, which primarily consists of instrumental music. Expanding the training dataset to include vocal music could enable the generation of music with human voice, offering a more comprehensive music generation system.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. 2023. Musiclm: Generating music from text. *CoRR*, abs/2301.11325.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. 2021. Wavegrad: Estimating gradients for waveform generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794.

Gaëtan Hadjeres and Léopold Crestel. 2020. Vector quantized contrastive predictive coding for template-based music generation. *arXiv preprint arXiv:2004.10120*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Havnø Frank, Jesse H. Engel, Quoc V. Le, William Chan, and Wei Han. 2023. Noise2music: Text-conditioned music generation with diffusion models. *CoRR*, abs/2302.03917.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 2021. Variational diffusion models. *CoRR*, abs/2107.00630.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre D'efossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *ArXiv*, abs/2209.15352.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *CoRR*, abs/1411.1784.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Pasini and Jan Schlüter. 2022. Musika! fast infinite waveform music generation. *CoRR*, abs/2208.08706.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. 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*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer.

Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. 2023. Moûsai: Text-to-music generation with long-context latent diffusion. *CoRR*, abs/2301.11757.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3483–3491.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Belinda Tzen and Maxim Raginsky. 2019. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *CoRR*, abs/1905.09883.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Shangda Wu and Maosong Sun. 2022. Exploring the efficacy of pre-trained checkpoints in text-to-music generation task. *CoRR*, abs/2211.11216.

Yixiao Zhang, Ziyu Wang, Dingsu Wang, and Gus Xia. 2020. Butter: A representation learning framework for bi-directional music-sentence retrieval and generation. In *Proceedings of the 1st workshop on nlp for music and audio (nlp4musa)*, pages 54–58.

Le Zhuo, Zhaokai Wang, Baisen Wang, Yue Liao, Stanley Peng, Chenxi Bao, Miao Lu, Xiaobo Li, and Si Liu. 2022. Video background music generation: Dataset, method and evaluation. *CoRR*, abs/2211.11248.

A Dataset

Examples of our collected dataset can be seen in Table 6.

B Implemention Details

We train the model for 580,000 steps using Adam optimizers with a learning rate of 4e-5 and a training batch size of 96. We save exponential moving averaged model weights with a decay rate of 0.995, except for the first 25 epochs.

C Music Tags Extraction

To obtain the music tags, we use the TF-IDF model to mine terms with higher frequency and importance from the dataset. Given a set of text Y, the basic assumption is that the texts contain various words or phrases related to music features such as instruments and genres. We aim to mine a tag set T from Y. We assume two rules to define a good music tag representing typical music features: 1) A certain amount of different music can be described with the tag for the model to learn the "text(tag)-to-music" mapping without loss of diversity; 2) A tag is worthless if it appears in the descriptions of too many pieces of music. For example, almost every piece of music can be described as "good listening"; thus, it should not be adopted as a music tag. Based on such rules, we leverage the TF-IDF model to mine the music tags. Because the language of our dataset is Chinese, we use jieba[5] to cut the sentences into terms. For a term w, we make statistics on the total dataset to obtain the TF $\mathrm{tf}(w)$ and the IDF $\mathrm{idf}(w)$, then the term score is obtained as $\mathrm{score}(w) = \mathrm{tf}(w) \cdot \mathrm{idf}(w)$. We reversely sort all

Table 5: Examples of free-form texts and corresponding music tags.

Text	Tags
聆听世界著名的钢琴曲简直是一种身心享受，我非常喜欢 Listening to the world famous piano music is simply a kind of physical and mental enjoyment, I like it very much	钢琴 piano
钢琴旋律的弦音，轻轻地、温柔地倾诉心中的遐想、心中的爱恋 The strings of the piano melody, gently and tenderly express the reverie and love in the heart	钢琴，轻轻，温柔，爱 piano, gentle, tender, love
提琴与钢琴合鸣的方式，在惆怅中吐露出淡淡的温柔气息 The ensemble of violin and piano reveals a touch of gentleness in melancholy	钢琴, 小提琴, 温柔, 惆怅 piano, violin, gentle, melancholic

Table 6: Examples of the adopted and abandoned tags

	Tags
Adopted	希望，生命，钢琴，小提琴，孤独，温柔，幸福，悲伤，游戏，电影 hope, life, piano, violin, lonely, gentle, happiness, sad, game, movie
Abandoned	音乐，喜欢，感觉，世界，好听，旋律，永远，音符，演奏，相信 music, like, feeling, world, good-listening, melody, forever, note, play, believe

[5]https://github.com/fxsjy/jieba

Title	Musician	Text
风的礼物 Gift of the Wind	西村由纪江 Yukie Nishimura	轻快的节奏，恰似都市丽人随风飘过的衣袂。放松的心情，片刻的愉快驱散的是工作的压力和紧张，沉浸其中吧，自己的心。 The brisk rhythm is like the clothes of urban beauties drifting in the wind. A relaxed mood, a moment of pleasure, dispels the pressure and tension of work. Immerse yourself, your own heart, in it.
九龙水之悦 Joy of the Kowloon Water	李志辉 Zhihui Li	聆听［九龙水之悦］卸下所有的苦恼，卸下所有的沉重，卸下所有的忧伤，还心灵一份纯净，还人生一份简单。 Listen to “The Joy of the Kowloon Water" to remove all the troubles, all the heaviness, and all the sorrows and restore the purity of the soul and the simplicity of life.
白云 Nuvole Bianche	鲁多维科·伊诺 Ludovico Einaudi	钢琴的更宁静，可大提琴的更多的是悠扬和深沉，也许是不同的演奏方式带来不同的音乐感受吧。 The piano is more serene, but the cello is more melodious and deep. Perhaps different playing methods bring different musical feelings.

Table 7: Examples of our Web Music with Text dataset.

Algorithm 1 Training

repeat

$x \sim p(x|y)$

$t \sim \text{Uniform}([0, 1])$

$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$v_t \leftarrow \alpha_t \epsilon - \sigma_t x$

Take gradient step on

$\nabla_\theta \|v_t - \hat{v}_\theta(\alpha_t x + \sigma_t \epsilon, t, y)\|^2$

until converged

the terms based on $\text{score}(w)$ and manually select 100 best music tags to obtain the ultimate music tag set T, which can represent the features of music such as instruments, music genres, and expressed emotions. Table 6 displays examples of the adopted and abandoned terms.

We use the mined music tags to condition the diffusion process. For a piece of music from the training data, we concatenate its corresponding music tags with a separator symbol “, ” to obtain a music tag sequence as the conditioning text to train the model.