JEN-1: Text-Guided Universal Music Generation with Omnidirectional Diffusion Models

Peike Li[1], Boyu Chen, Yao Yao,
Yikai Wang, Allen Wang, Alex Wang[1]
Futureverse AI Research
peike.li@yahoo.com, alex.wang@futureverse.com

Abstract—**Music generation has attracted growing interest with the advancement of deep generative models. However, generating music conditioned on textual descriptions, known as text-to-music, remains challenging due to the complexity of musical structures and high sampling rate requirements. Despite the task's significance, prevailing generative models exhibit limitations in music quality, computational efficiency, and generalization ability. This paper introduces JEN-1, a universal high-fidelity model for text-to-music generation. JEN-1 is a diffusion model incorporating both autoregressive and non-autoregressive training in an end-to-end manner, enabling up to 48kHz high-fidelity stereo music generation. Through multi-task in-context learning, JEN-1 performs various generation tasks including text-guided music generation, music inpainting, and continuation. Evaluations demonstrate JEN-1's superior performance over state-of-the-art methods in text-music alignment and music quality while maintaining computational efficiency. Our demo pages are available at https://jenmusic.ai/audio-demos**

Index Terms—**generative models, text-to-music, music generation, non-autoregressive, diffusion models**

I. INTRODUCTION

Music, as an artistic expression comprising harmony, melody and rhythm, holds great cultural significance and appeal to humans. Recent years have witnessed remarkable progress in music generation with the rise of deep generative models [1]–[3]. However, generating high-fidelity and realistic music still poses unique challenges compared to other modalities. Firstly, music utilizes the full frequency spectrum, requiring high sampling rates like 44.1kHz stereo to capture the intricacies. This is in contrast to speech generation which focuses on linguistic content and uses lower sampling rates (*e.g.*, 16kHz). Secondly, the blend of multiple instruments and the arrangement of melodies and harmonies result in highly complex structures. With humans being sensitive to musical dissonance, music generation allows little room for imperfections. Most critically, controllability over attributes like key, genre and melody is indispensable for creators to realize their artistic vision.

The multi-modal intersection of text and music, known as text-to-music generation, offers valuable capabilities to bridge free-form textual descriptions and musical compositions. A variety of works have been done towards text-to-music generation. Regarding how they encode the high complexity of input music signals, typical methods are generally categorized

into, (1) directly adopting waveform signals with the help of discrete encodings to handle the complexity burden [3], [4], or (2) converting signals to spectrum formats that allow continuous encodings, which, inevitably incur fidelity losses during the spectrum conversion process [1], [5]. Apart from these works, Noise2Music [6] tries to leverage waveform signals with continuous encodings yet the fidelity of music representation is bottle-necked to only 3.2kHz, which is still limited even after adding a cascade model and a super-resolution stage hierarchically.

A systematic comparison of existing methods is provided in Table I, where some of the approaches operate on spectrogram representations, incurring fidelity loss from audio conversion [1], [5]. Others employ inefficient autoregressive designs to sequentially predict each music token [3], or adopt multiple models to cascade-generate the music, resulting in low computational efficiency [4], [6]. More restrictively, the training objectives of existing methods are confined to a single task, lacking the versatility to conduct multiple and general music generation and editing tasks.

In this paper, we propose **JEN-1**, a universal text-to-music generation model combining quality, controllability, and efficiency. JEN-1 leverages a hybrid autoregressive (AR) and non-autoregressive (NAR) structure, with a novel omnidirectional design that potentially enjoys the benefits of both generation efficiency and quality. By the combination of AR and NAR structural designs, for the first time, JEN-1 unifies the learning process of multiple tasks including text-to-music, music inpainting, and music continuation into one single model. Another hallmark of JEN-1 is the ability to encode raw waveform data formats, *e.g.*, enabling the generation of high-fidelity 48kHz stereo audios, which is realized by our proposed masked noise-robust autoencoder with specific designs for continuous embedding representation and latent embedding normalization.

We extensively evaluate JEN-1 against state-of-the-art baselines across objective metrics and human evaluations. Results indicate that JEN-1 produces music of perceptually higher quality compared to the current best methods (85.7 *vs.* 83.8). Ablations validate the efficacy of each technical component and also demonstrate that JEN-1 benefits greatly from multi-task joint training. More importantly, human judges confirm JEN-1 generates music highly aligned with text prompts in a melodically and harmonically pleasing fashion.

---

[1]Corresponding authors.

TABLE I
COMPARISON BETWEEN STATE-OF-THE-ART MUSIC GENERATIVE MODELS.

	Feature	MusicLM	MusicGen	AudioLDM	Noise2Music	**JEN-1 (Ours)**
Data	high sample rate	✗	✗	✗	✗	✓
	2-channel stereo	✗	✗	✗	✗	✓
	waveform	✓	✓	✗	✓	✓
Model	autoregressive	✓	✓	✗	✗	✓
	non-autoregressive	✗	✗	✓	✓	✓
	non-cascade model	✗	✓	✓	✗	✓
Task	single-task training	✓	✓	✓	✓	✓
	multi-task training	✗	✗	✗	✗	✓

In summary, the key contributions of this work are:

1) We propose JEN-1 as a universal framework to the challenging text-to-music generation task. JEN-1 utilizes an extremely efficient approach by directly modeling waveforms and avoids the conversion loss associated with spectrograms. It incorporates a masked autoencoder and diffusion model, yielding high-quality music at a 48kHz sampling rate.

2) JEN-1 integrates both autoregressive and non-autoregressive diffusion modes in an end-to-end manner to improve sequential dependency and enhance sequence generation concurrently, enabling music generation, music continuation, and music inpainting within one single non-cascade model.

3) We conduct comprehensive evaluations, both objective and involving human judgment, to thoroughly assess the crucial design choices underlying our method. Results demonstrate that JEN-1 generates melodically aligned music that adheres to textual descriptions while maintaining high fidelity.

II. RELATED WORK

In this section, we present a review of the extant literature within the domain of music generation. We undertake a comparative analysis of three primary paradigmatic distinctions: single-task *vs.* multi-task training, waveform *vs.* spectrum-based methods, and autoregressive *vs.* non-autoregressive generative modes.

Single-task *vs.* Multi-task. Several prior symbolic music generation methods [7], [8] have primarily produced basic MIDI-style compositions, characterized by their simplicity and lack of realism. In contrast, recent methodologies [1], [3], [6] have employed text descriptions as conditioning inputs to facilitate the direct generation of high-fidelity raw musical compositions. Nevertheless, existing approaches are often constrained by narrow generation objectives, limiting their applicability to a singular task and impeding their adaptability for diverse music generation and editing tasks. In contrast, JEN-1 introduces a universal framework designed to tackle the intricate text-to-music generation challenge. This framework streamlines the learning process by encompassing multiple tasks within a singular mode, thus encompassing text-to-music conversion, music inpainting, and music continuation.

Waveform *vs.* Spectrum. The fundamental processes of audio feature extraction and representation learning [9]–[12] play a pivotal role in music generation and can be classified into two main directions. One of these methods entails the transformation of the audio waveform into a mel-spectrogram [1], [5], which is then processed as images. However, this transformation from waveform to spectrogram necessitates a shift from continuous encodings to spectral formats, a step that inevitably introduces fidelity losses in the audio data. Conversely, the other approaches [3], [6] involve the direct utilization of raw waveform-based audio data. Some variations of this approach go a step further by converting the audio into discrete tokens [3], [4] for modeling purposes, also resulting in a compromise in audio quality. In this study, JEN-1 adopts a strategy that preserves the raw waveform data in a continuous space format, facilitating the generation of high-fidelity stereo audio at a sampling rate of 48kHz.

Autoregressive *vs.* Non-autoregressive. Music generation draws inspiration from both natural language processing (NLP) and computer vision (CV), incorporating both autoregressive (AR) and non-autoregressive (NAR) models. AR models, such as PerceiverAR [13], AudioGen [2], MusicLM [3], and Jukebox [14], predict audio tokens sequentially based on prior context. However, their processing speed limitations hinder their utility in music in-painting tasks. Conversely, NAR models generate multiple tokens concurrently, offering faster processing and improved generative performance. Modern NAR models, such as StableDiffusion [15], excel in image generation, while diffusion models [16] have gained attention for their NAR generative capabilities in music generation tasks. These models progressively refine random noise into latent representations, enabling efficient synthesis of high-quality audio. Some approaches, including Make-An-Audio [17], Noise2Music [6], AudioLDM [1], and TANGO [5], extend latent diffusion models (LDM) [15] and demonstrate significant speed advantages in music generation tasks. In this study, JEN-1 pioneers a hybrid AR and NAR structure to enhance sequential dependency and improve concurrent sequence generation.

III. PRELIMINARY

A. Conditional Generative Models

In the field of content synthesis, the implementation of conditional generative models often involves applying either autoregressive (AR) [3], [4] or non-autoregressive (NAR) [1], [5] paradigms. The inherent structure of language, where each word functions as a distinct token and sentences are sequentially constructed from these tokens, makes the AR paradigm a more natural choice for language modeling. Thus, in the domain of Natural Language Processing (NLP), transformer-based models, *e.g.*, GPT series, have emerged as the prevailing approach for text generation tasks. AR methods [3], [4] rely on predicting future tokens based on visible history tokens. The likelihood is represented by:

$$p_{\text{AR}}(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{N} p\left(\boldsymbol{y}_i \mid \boldsymbol{y}_{1:i-1}; \boldsymbol{x}\right), \tag{1}$$

where $\boldsymbol{y}_i$ represents the i-th token in sequence $\boldsymbol{y}$.

Conversely, for the image generation task where images have no explicit time series structure and images typically occupy continuous space, employing an NAR approach is deemed more suitable. Notably, the NAR approach, such as stable diffusion, has emerged as the dominant method for addressing image generation tasks. NAR approaches assume conditional independence among latent embeddings and generate them uniformly without distinction during prediction. This results in a likelihood expressed as:

$$p_{\text{NAR}}(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{N} p\left(\boldsymbol{y}_i \mid \boldsymbol{x}\right). \tag{2}$$

Although the parallel generation approach of NAR offers a notable speed advantage, it falls short in terms of capturing long-term consistency.

In this work, we argue that audio data can be regarded as a hybrid form of data. It exhibits characteristics akin to images, as it resides within a continuous space that enables the modeling of high-quality music. Additionally, audio shares similarities with text in its nature as a time-series data. Consequently, we propose a novel approach in our JEN-1 design, which entails the amalgamation of both the auto-regressive and non-autoregressive modes into a cohesive omnidirectional diffusion model.

B. Diffusion Models for Audio Generation

Diffusion models [16] constitute probabilistic models explicitly developed for the purpose of learning a data distribution $p(\boldsymbol{x})$. The overall learning of diffusion models involves a forward *diffusion* process and a gradual *denoising* process, each consisting of a sequence of T steps that act as a Markov Chain. In the forward diffusion process, a fixed linear Gaussian model is employed to gradually perturb the initial random vari-

able $\boldsymbol{z}_0$ until it converges to the standard Gaussian distribution. This process can be formally articulated as follows,

$$q\left(\boldsymbol{z}_t \mid \boldsymbol{z}_0; \boldsymbol{x}\right) = \mathcal{N}\left(\boldsymbol{z}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0, \left(1 - \bar{\alpha}_t\right)\mathbf{I}\right),$$
$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i, \tag{3}$$

where α_i is a coefficient that monotonically decreases with timestep t, and $\boldsymbol{z}_t$ is the latent state at timestep t. The reverse process is to initiate from standard Gaussian noise and progressively utilize the denoising transition $p_{\boldsymbol{\theta}}\left(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t; \boldsymbol{x}\right)$ for generation,

$$p_{\boldsymbol{\theta}}\left(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t; \boldsymbol{x}\right) = \mathcal{N}\left(\boldsymbol{z}_{t-1}; \mu_{\boldsymbol{\theta}}\left(\boldsymbol{z}_t, t; \boldsymbol{x}\right), \Sigma_{\boldsymbol{\theta}}\left(\boldsymbol{z}_t, t; \boldsymbol{x}\right)\right), \tag{4}$$

where the mean $\mu_{\boldsymbol{\theta}}$ and variance $\Sigma_{\boldsymbol{\theta}}$ are learned from the model parameterized by θ. We use predefined variance without trainable parameters following [1], [15]. After simply expanding and re-parameterizing, our training objective of the conditional diffusion model can be denoted as,

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_{\boldsymbol{\theta}}\left(\boldsymbol{z}_t, t\right)\|_2^2\right], \tag{5}$$

where t is uniformly sampled from $\{1, ..., T\}$, ϵ is the ground truth of the sampled noise, and $\epsilon_{\theta}(\cdot)$ is the noise predicted by the diffusion model.

The conventional diffusion model is characterized as a non-autoregressive model, which poses challenges in effectively capturing sequential dependencies in music flow. To address this limitation, we propose the joint omnidirectional diffusion model JEN-1, an integrated framework that leverages both unidirectional and bidirectional training. These adaptations allow for precise control over the contextual information used to condition predictions, enhancing the model's ability to capture sequential dependencies in music data.

IV. METHOD

In this research paper, we propose a novel model called JEN-1, which utilizes an omnidirectional 1D diffusion model. JEN-1 combines bidirectional and unidirectional modes, offering a unified approach for universal music generation conditioned on either text or music inputs. The model operates in a noise-robust latent embedding space obtained from a masked audio autoencoder, enabling high-fidelity reconstruction from latent embeddings with a low frame rate(§ IV-A). In contrast to prior generation models that use discrete tokens or involve multiple cascaded stages, JEN-1 introduces a unique modeling framework capable of generating continuous, high-fidelity music using one single model. JEN-1 effectively utilizes both autoregressive training to improve sequential dependency and non-autoregressive training to enhance sequence generation concurrently (§ IV-B). By employing in-context learning and multi-task learning, one of the significant advantages of JEN-1 is its support for conditional generation based on either text or melody, enhancing its adaptability to various creative scenarios (§ IV-C). This flexibility allows the model to be applied to different music generation tasks, making it a versatile and powerful tool for music composition and production.
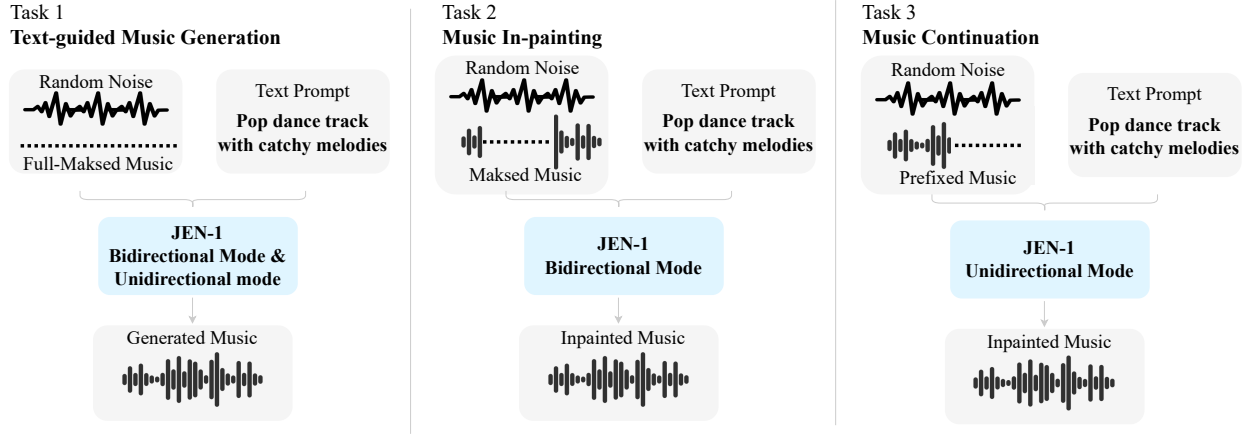
Fig. 1. Illustration of the JEN-1 multi-task training strategy, including the text-guided music generation task, the music inpainting task, and the music continuation task. JEN-1 achieves the in-context learning task generalization by concatenating the noise and masked audio in a channel-wise manner. JEN-1 integrates both the bidirectional mode to gather comprehensive context and the unidirectional mode to capture sequential dependency.

A. Masked Autoencoder for High Fidelity Latent Representation Learning

High Fidelity Neural Audio Latent Representation. To facilitate the training on limited computational resources without compromising quality and fidelity, our approach JEN-1 employs a high-fidelity audio autoencoder $\mathcal{E}$ to compress original audio into latent representations z. Formally, given a two-channel stereo audio $x \in \mathbb{R}^{L \times 2}$, the encoder $\mathcal{E}$ encodes x into a latent representation $z = \mathcal{E}(x)$, where $z \in \mathbb{R}^{L/h \times c}$. L is the sequence length of given music, h is the hop size and c is the dimension of latent embedding. While the decoder reconstructs the audio $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$ from the latent representation. Our audio compression model is inspired and modified based on previous work [18], [19], which consists of an autoencoder trained by a combination of a reconstruction loss over both time and frequency domains and a patch-based adversarial objective operating at different resolutions. This ensures that the audio reconstructions are confined to the original audio manifold by enforcing local realism and avoids muffled effects introduced by relying solely on sample-space losses with L1 or L2 objectives. Unlike prior endeavors [18], [19] that employ a quantization layer to produce the discrete codes, our model directly extracts the continuous embeddings without any quality-reducing loss due to quantization. This utilization of powerful autoencoder representations enables us to achieve a nearly optimal balance between complexity reduction and high-frequency detail preservation, leading to a significant improvement in music fidelity.

Noise-robust Masked Autoencoder. To further enhance the robustness of decoder $\mathcal{D}$, we propose a masking strategy, which effectively reduces noises and mitigates artifacts, yielding superior-quality audio reconstruction. In our training procedure, we adopt a specific technique wherein $p = 5\%$ of the intermediate latent embeddings are randomly masked before feeding into the decoder. By doing so, we enable the decoder to acquire proficiency in reconstructing superior-quality data even when exposed to corrupted inputs. We train the autoencoder on

---

Algorithm 1 Normalizing Latent Embedding Space

Input: Existing latent embeddings $\{z_i\}_{i=1}^{N}$ and reduced dimension k

1: compute μ and Σ of $\{z_i\}_{i=1}^{N}$
2: compute $U, \Lambda, U^T = \text{SVD}(\Sigma)$
3: compute $W = (U\sqrt{\Lambda^{-1}})[:,:k]$
4: $\tilde{z}_i = (z_i - \mu)W$

Output: Normalized latent embeddings $\{\tilde{z}_i\}_{i=1}^{N}$

48kHz stereophonic audios with large batch size and employ an exponential moving average to aggregate the weights. As a result of these enhancements, the performance of our audio autoencoder surpasses that of the original model in all evaluated reconstruction metrics. Consequently, we adopt this audio autoencoder for all of our subsequent experiments.

Normalizing Latent Embedding Space. To avoid arbitrarily scaled latent spaces, [15] found it is crucial to achieve better performance by estimating the component-wise variance and re-scale the latent z to have a unit standard deviation. In contrast to previous approaches that only estimate the component-wise variance, JEN-1 employs a straightforward yet effective post-processing technique to address the challenge of anisotropy in latent embeddings as shown in Algorithm 1. Specifically, we channel-wisely perform zero-mean normalization on the latent embedding, and then transform the covariance matrix to the identity matrix via the Singular Value Decomposition (SVD) algorithm. We implement a batch-incremental equivalent algorithm to calculate these transformation statistics. Additionally, we incorporate a dimension reduction strategy to keep 90% most important channels (reduced dimension k in Algorithm 1) to enhance the whitening process further and improve the overall effectiveness of our approach.

B. omnidirectional Latent Diffusion Models

In some prior approaches [1], [5], time-frequency conversion techniques, such as mel-spectrogram, have been employed
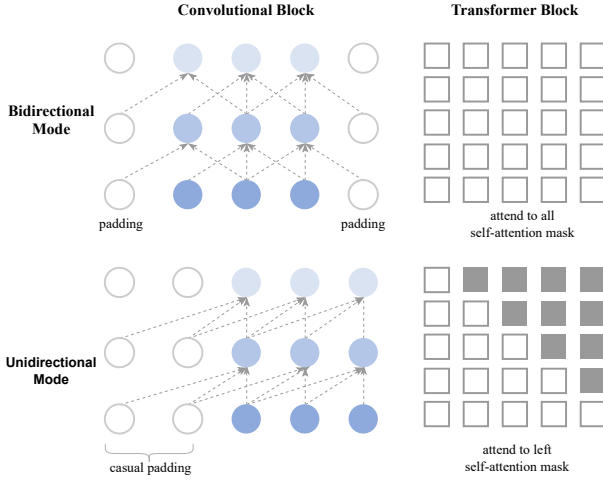
Fig. 2. Illustration of bidirectional mode and unidirectional mode for convolutional block and transformer block. In the unidirectional mode, we use causal padding in the convolutional block and attend the self-attention mask only to the left context in the transformer block.

for transforming the audio generation into an image generation problem. Nevertheless, we contend that this conversion from raw audio data to mel-spectrogram inevitably leads to a significant reduction in quality. To address this concern, JEN-1 directly leverages a temporal 1D efficient U-Net. This modified version of the Efficient U-Net [20] allows us to effectively model the waveform and implement the required blocks in the diffusion model. The U-Net model's architecture comprises cascading down-sampling and up-sampling blocks interconnected via residual connections. Each down/up-sampling block consists of a down/upsampling layer, followed by a set of blocks that involve 1D temporal convolutional layers, and self/cross-attention layers. Both the stacked input and output are represented as latent sequences of length L, while the diffusion time t is encoded as a single-time embedding vector that interacts with the model via the aforementioned combined layers within the down and up-sampling blocks. In the context of the U-Net model, the input consists of the noisy sample denoted as x_t, which is stacked with additional conditional information. The resulting output corresponds to the noise prediction ϵ during the diffusion process.

Task Generalization via In-context Learning. To better achieve the goal of multi-task versatility, we propose a novel omnidirectional latent diffusion model without explicitly changing the U-Net architecture. As shown in Figure 1, JEN-1 formulates various music generation tasks as text-guided in-context learning tasks. The common goal of these in-context learning tasks is to produce diverse and realistic music that is coherent with the context music and has the correct style described by the text. For in-context learning objectives, *e.g.*, music inpainting task, and music continuation task, additional masked music information, which the model is conditioned upon, can be extracted into latent embedding and stacked as additional channels in the input. More precisely, apart from the original latent channels, the U-Net block has 129 additional

input channels (128 for the encoded masked audio and 1 for the mask itself).

From Bidirectional Mode to Unidirectional Mode. To account for the inherent sequential characteristic of music, JEN-1 integrates the unidirectional diffusion mode by ensuring that the generation of latent on the right depends on the generated ones on the left, a mechanism achieved through employing a unidirectional self-attention mask and a causal padding mode in convolutional blocks. In general, the architecture of the omnidirectional diffusion model enables various input pathways, facilitating the integration of different types of data into the model, resulting in versatile and powerful capabilities for noise prediction and diffusion modeling. During training, JEN-1 could switch between a unidirectional mode and a bidirectional model without changing the architecture of the model. The parameter weight is shared for different learning objectives. As illustrated in Figure 2, JEN-1 could switch into the unidirectional (autoregressive) mode, *i.e.*, the output variable depends only on its own previous values. We employ *causal padding* [21] in all 1D convolutional layers, padding with zeros in the front so that we can also predict the values of early time steps in the frame. In addition, we employ a triangular attention mask following [22], by padding and masking future tokens in the input received by the self-attention blocks.

C. Unified Music Multi-task Training

In contrast to prior methods that solely rely on a single text-guided learning objective, our proposed framework, JEN-1, adopts a novel approach by simultaneously incorporating multiple generative learning objectives while sharing common parameters. As depicted in Figure 1, the training process encompasses three distinct music generation tasks: bidirectional text-guided music generation, bidirectional music inpainting, and unidirectional music continuation. The utilization of multi-task training is a notable aspect of our approach, allowing for a cohesive and unified training procedure across all desired music generation tasks. This approach enhances the model's ability to generalize across tasks, while also improving the handling of music sequential dependencies and the concurrent generation of sequences.

Text-guided Music Generation Task. In this task, we employ both the bidirectional and unidirectional modes. The bidirectional model allows all latent embeddings to attend to one another during the denoising process, thereby enabling the encoding of comprehensive contextual information from both preceding and succeeding directions. On the other hand, the unidirectional model restricts all latent embeddings to attend solely to their previous time counterparts, which facilitates the learning of temporal dependencies in music data. Moreover, for the purpose of preserving task consistency within the framework of U-Net stacked inputs, we concatenate a full-size mask alongside all-empty masked audio as the additional condition.

Music inpainting Task. In the domain of audio editing, inpainting denotes the process of restoring missing segments

METHODS	QUANTITATIVE			QUALITATIVE		EFFICIENCY
	FAD↓	KL ↓	CLAP↑	T2M-QLT ↑	T2M-ALI ↑	PARAMETERS ↓
Riffusion	14.8	2.06	0.19	72.1	72.2	890M
Mousai [23]	7.5	1.59	0.23	76.3	71.9	857M
MusicLM†‡ [3]	4.0	-	-	81.7	82.0	860M
AudioLDM [1]	2.3	1.35	0.31	81.9	71.8	739M
Noise2Music‡ [6]	2.1	-	-	-	-	1.3B
MusicGen [4]	3.8	**1.22**	0.31	83.8	79.5	3.3B
JEN-1 (Ours)	**2.0**	1.29	**0.33**	**85.7**	**82.8**	**726M**

within the music. This restorative technique is predominantly employed to reconstruct corrupted audio from the past, as well as to eliminate undesired elements like noise and watermarks from musical compositions. In this task, we adopt the bidirectional mode in JEN-1. During the training phase, our approach involves simulating the music inpainting process by randomly generating audio masks with mask ratios ranging from 20% to 80%. These masks are then utilized to obtain the corresponding masked audio, which serves as the conditional in-context learning inputs within the U-Net model.

Music Continuation Task. We demonstrate that the proposed JEN-1 model facilitates both music inpainting (interpolation) and music continuation (extrapolation) by employing the novel omnidirectional diffusion model. The conventional diffusion model, due to its non-autoregressive nature, has demonstrated suboptimal performance in previous studies [3], [24]. This limitation has impeded its successful application in audio continuation tasks. To address this issue, we adopt the unidirectional mode in our music continuation task, ensuring that the predicted latent embeddings exclusively attend to their leftward context within the target segment. Similarly, we simulate the music continuation process through the random generation of exclusive right-only masks. These masks are generated with varying ratios spanning from 20% to 80%.

The overall training objective is the sum of the different types of JEN-1 tasks described above. Specifically, within a training batch, for 1/3 of the data we use the bidirectional text-guided generation objective, for 1/3 of the data we employ the bidirectional music inpainting objective, and the unidirectional music continuation objective is calculated with the rest of 1/3 data. To gain a more comprehensive overview of JEN-1, we kindly direct your attention to our demo page.

V. EXPERIMENT

Implementation Details. For the masked music autoencoder, we used a hop size of 320, resulting in 125Hz latent sequences for encoding 48kHz music audio. The dimension of latent embedding is 128. We randomly mask 5% of the latent embedding during training to achieve a noise-tolerant decoder. We employ FLAN-T5 [25], an instruct-based large language model to provide superior text embedding extraction.

For the omnidirectional diffusion model, we implemented a 1D conditional UNet latent diffusion model inspired by Stable Diffusion [15]. we set the intermediate cross-attention dimension to 1024, resulting in 726M parameters. During the multi-task training, we evenly allocate 1/3 of a batch to each training task. In addition, we applied the classifier-free guidance [26] to improve the correspondence between samples and text conditions. During training, the cross-attention layer is randomly replaced by self-attention with a probability of 0.2. We train our JEN-1 models on 8 A100 GPUs for 200k steps with the AdamW optimizer [27], a linear-decayed learning rate starting from $3e^{-5}$ a total batch size of 512 examples, $\beta_1 = 0.9$, $\beta_2 = 0.95$, a decoupled weight decay of 0.1, and gradient clipping of 1.0. In the diffusion forward process, we use the DDPM [16] sampler with $N = 1000$ steps, and a linear noise schedule from $\beta_1 = 0.0015$ to $\beta_N = 0.0195$ is used. For classifier-free guidance, a guidance scale w of 2.0 is used. While it is feasible to employ an advanced sampling scheduler to markedly enhance inference speed, to maintain comparison fairness, we adhere to the approach outlined in [1] during the inference process. Specifically, we utilize the DDIM [28] sampler with 200 sampling steps.

Datasets. We use total 15k samples of high-quality licensed music from Pond5 to train JEN-1. All music data consist of full-length music sampled at 48kHz with metadata composed of a rich textual description and additional tags information, *e.g.*, genre, instrument, mood/theme tags, *etc*. The proposed method is evaluated using the MusicCaps [3] benchmark, which consists of 5.5K expert-prepared music samples, each lasting ten seconds, and a genre-balanced subset containing 1K samples. To maintain fair comparison, objective metrics are reported on the unbalanced set, while qualitative evaluations and ablation studies are conducted on examples randomly sampled from the genre-balanced set.

Evaluation Metrics. In our quantitative evaluation, we employ both objective and subjective metrics following [4]. Objective evaluation encompasses three metrics: Fréchet Audio Distance (FAD) [29], Kullback-Leibler Divergence (KL) [30], and CLAP score (CLAP) [31]. FAD gauges audio plausibility, with a lower score indicating higher plausibility. KL-divergence

TABLE III
ABLATION STUDIES. FROM THE BASELINE CONFIGURATION, WE INCREMENTALLY MODIFY THE JEN-1 CONFIGURATION TO INVESTIGATE THE EFFECT OF EACH COMPONENT.

CONFIGURATION	QUANTITATIVE			QUALITATIVE	
	FAD↓	KL ↓	CLAP↑	T2M-QLT ↑	T2M-ALI ↑
baseline	3.1	1.35	0.31	80.1	78.3
+ noise-robust autoencoder	3.1	1.36	0.31	80.5	78.1
+ latent space normalization	2.6	1.34	0.32	81.9	78.9
+ auto-regressive mode	2.5	1.33	0.33	82.9	79.5
+ music in-painting task	2.2	**1.28**	0.32	83.8	80.1
+ music continuation task	**2.0**	1.29	**0.33**	**85.7**	**82.8**

measures the similarity between generated and reference music in terms of label probabilities, utilizing a state-of-the-art audio classifier trained on AudioSet [32]. A lower KL score suggests greater conceptual similarity. Additionally, the CLAP score quantifies audio-text alignment, using the official pre-trained CLAP model. For qualitative assessments, we follow the experimental design outlined in Copet et al. [4]. Two aspects of the generated music are evaluated qualitatively: text-to-music quality (T2M-QLT) and alignment to the text input (T2M-ALI). Human raters assigned perceptual quality ratings on a scale of 1 to 100 in the text-to-music quality test. In the text-to-music alignment test, raters assessed the alignment between audio and text using the same rating scale.

A. Comparison with State-of-the-arts

As shown in Table II, we compare the performance of JEN-1 with other state-of-the-art methods, including Riffusion [33], and Mousai [23], MusicLM [3], MusicGen [4], Noise2Music [6]. These competing approaches were all trained on large-scale music datasets and demonstrated state-of-the-art music synthesis ability given diverse text prompts. To ensure a fair comparison, we evaluate the performance on the MusicCaps test set from both quantitative and qualitative aspects. Since the implementation is not publicly available, we utilize the MusicLM public API for our tests. And for Noise2Music, we only report the FAD score as mentioned in their original paper. Experimental results demonstrate that JEN-1 outperforms other competing baselines concerning both text-to-music quality and text-to-music alignment. Specifically, JEN-1 exhibits superior performance in terms of FAD and CLAP scores, outperforming the second-highest method Noise2Music and MusicGen by a large margin. Regarding the human qualitative assessments, JEN-1 consistently achieves the best T2M-QLT and T2M-ALI scores. It is noteworthy that our JEN-1 is more computationally efficient with only 22.6% of MusicGEN (726M *vs.* 3.3B parameters) and 57.7% of Noise2Music (726M *vs.* 1.3B parameters).

B. Performance Analysis

Ablation Studies. To assess the effects of the omnidirectional diffusion model, we compare the different configurations, including the effect of model configuration and the effect

of different multi-task objectives. All ablations are conducted on 1K genre-balanced samples, randomly selected from the held-out evaluation set. As illustrated in Table III, the results demonstrate that i) the music generation quality and performance are substantially enhanced by our specifically designed noise-robust autoencoder and normalized latent space; ii) JEN-1 incorporates the auto-regressive mode greatly benefiting the temporal consistency of generated music, leading to better music quality; iii) our proposed multi-task learning objectives, *i.e.*, text-guided music generation, music inpainting, and music-continuation, improve task generalization and consistently achieve better performance; iv) all these dedicated designs together lead to high-fidelity music generation without introducing much extra training cost.

Generation Diversity. Compared to transformer-based generation methods, diffusion models are notable for their generation diversity. To further investigate JEN-1's generation diversity and credibility, we provide identical textual prompts, such as descriptions involving general genres or instruments, to generate multiple different samples. As demonstrated on our demo page, JEN-1 showcases impressive diversity in its generation outputs while maintaining a consistently high level of quality.

Generalization and Controllability. Despite being trained with paired texts and music samples in a supervised learning manner, our method, JEN-1, demonstrates noteworthy zero-shot generation capability and effective controllability. Notwithstanding the challenges associated with generating high-quality audio from out-of-distribution prompts, JEN-1 still demonstrates its proficiency in producing compelling music samples. On our demo page, we present examples of creative zero-shot prompts, showcasing the model's successful generation of satisfactory quality music. Furthermore, we present generation examples as evidence of JEN-1's proficiency in capturing music-related semantics. Notably, our demo indicates that the generated music adequately reflects music concepts such as the genre, instrument, mood, speed, *etc.*.

VI. CONCLUSION

In this work, we propose JEN-1, a text-to-music framework that directly models waveforms and integrates autoregressive

and non-autoregressive training. Through multi-task objectives, JEN-1 achieves high-quality music generation from text descriptions. Extensive evaluations show JEN-1's superiority in quality, diversity, and controllability over strong baselines. This research advances the state-of-the-art in controllable text-to-music generation. Future directions include exploring hierarchical multi-stem music generation and external knowledge incorporation. We believe high-quality text-to-music generation will empower new creative workflows and reshape music composition and appreciation. As the field matures from research to practical applications, it bears great potential to augment human creativity.

REFERENCES

[1] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[2] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

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

[4] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *arXiv preprint arXiv:2306.05284*, 2023.

[5] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

[6] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank *et al.*, "Noise2music: Text-conditioned music generation with diffusion models," *arXiv preprint arXiv:2302.03917*, 2023.

[7] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, "Midinet: A convolutional generative adversarial network for symbolic-domain music generation," *arXiv preprint arXiv:1703.10847*, 2017.

[8] S. Sulun, M. E. Davies, and P. Viana, "Symbolic music generation conditioned on continuous-valued emotions," *IEEE Access*, vol. 10, pp. 44 617–44 626, 2022.

[9] P. Li, X. Yu, and Y. Yang, "Super-resolving cross-domain face miniatures by peeking at one-shot exemplar," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4469–4479.

[10] P. Li, Y. Wei, and Y. Yang, "Consistent structural relation learning for zero-shot segmentation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 317–10 327, 2020.

[11] X. Pan, P. Li, Z. Yang, H. Zhou, C. Zhou, H. Yang, J. Zhou, and Y. Yang, "In-n-out generative learning for dense unsupervised video segmentation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1819–1827.

[12] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3260–3271, 2020.

[13] C. Hawthorne, A. Jaegle, C. Cangea, S. Borgeaud, C. Nash, M. Malinowski, S. Dieleman, O. Vinyals, M. Botvinick, I. Simon *et al.*, "General-purpose, long-context autoregressive modeling with perceiver ar," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8535–8558.

[14] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[17] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," *arXiv preprint arXiv:2301.12661*, 2023.

[18] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[19] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[21] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] F. Schneider, Z. Jin, and B. Schölkopf, "Mo\ˆ usai: Text-to-music generation with long-context latent diffusion," *arXiv preprint arXiv:2301.11757*, 2023.

[24] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: a language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[25] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[26] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[28] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[29] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms." in *INTERSPEECH*, 2019, pp. 2350–2354.

[30] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[31] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[32] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[33] S. Forsgren and H. Martiros, "Riffusion - Stable diffusion for real-time music generation," 2022. [Online]. Available: https://riffusion.com/about