JEN-1 Composer: A Unified Framework for High-Fidelity Multi-Track Music Generation

Yao Yao[*] **Peike Patrick Li**[*] **Boyu Chen** **Alex Wang**
Futureverse, AI Innovation
{yao.yao, patrick.li, alex.wang}@futureverse.com

Abstract

With rapid advances in generative artificial intelligence, the text-to-music synthesis task has emerged as a promising direction for music generation from scratch. However, finer-grained control over multi-track generation remains an open challenge. Existing models exhibit strong raw generation capability but lack the flexibility to compose separate tracks and combine them in a controllable manner, differing from typical workflows of human composers. To address this issue, we propose JEN-1 Composer, a unified framework to efficiently model marginal, conditional, and joint distributions over multi-track music via a single model. JEN-1 Composer framework exhibits the capacity to seamlessly incorporate any diffusion-based music generation system, *e.g.* Jen-1, enhancing its capacity for versatile multi-track music generation. We introduce a curriculum training strategy aimed at incrementally instructing the model in the transition from single-track generation to the flexible generation of multi-track combinations. During the inference, users have the ability to iteratively produce and choose music tracks that meet their preferences, subsequently creating an entire musical composition incrementally following the proposed Human-AI co-composition workflow. Quantitative and qualitative assessments demonstrate state-of-the-art performance in controllable and high-fidelity multi-track music synthesis. The proposed JEN-1 Composer represents a significant advance toward interactive AI-facilitated music creation and composition. Demos will be available at https://www.jenmusic.ai/audio-demos.

"Composing is like driving down a foggy road toward a house. Slowly you see more details of the house, the color of the slates and bricks, the shape of the windows."

– Benjamin Britten

1 Introduction

With the rapid development of generative modeling, AI-driven music generation has become an emerging task that creates value for both research communities and the music industry. Pioneering works like Music Transformer (Huang et al., 2018) and MuseNet (Payne, 2019) operated on symbolic representations (Engel et al., 2017). Although capable of conditioning on textual description, their generated MIDI-style outputs tend to heavily depend on pre-defined virtual synthesizers, resulting in an unrealistic audio quality and limited diversity. More recent text-to-music approaches like MusicGen (Copet et al., 2023), MusicLM (Agostinelli et al., 2023), and Jen-1 (Li et al., 2023) have streamlined the procedure by by directly creating authentic audio waveforms based on textual prompts. This advancement enhances versatility and diversity without necessitating a deep understanding of music theory. Nonetheless, the results they produce consist of composite mixes rather than individual tracks (*e.g.*, bass, drum, instrument, melody tracks), limiting fine-grained control in comparison to the creative processes employed by human composers. Additionally, their choice of instruments and musical styles is influenced by the data on which they were trained, occasionally leading to unconventional combinations.

---

[*]equal contribution

The advent of multi-track recording technology has ushered in a new era of musical creativity, enabling composers to delve into intricate harmonies, melodies, and rhythms that go beyond what can be achieved with individual instruments (Zhu et al., 2020). Digital audio workstations provide artists with the means to expand their musical ideas without being constrained by temporal or spatial limitations. The wide range of available timbres grants composers greater freedom to explore their creative concepts. The practice of composing music one track at a time aligns well with the real-world workflows of musicians and producers. This approach allows for the iterative refinement of specific tracks, taking into consideration the impact of other tracks, thereby facilitating collaboration between humans and artificial intelligence (Frid et al., 2020). Nonetheless, creating separate models for diverse combinations of tracks comes with a prohibitively high cost. Our objective is to combine the flexibility of text-to-music generation with the control offered by multi-track modeling, in order to harmonize with versatile creative workflows.

To this end, we develop a unified generative framework, namely JEN-1 Composer, to jointly model the marginal, conditional, and joint distributions over multi-track music using one single model. By extending off-shelf text-to-music diffusion models with minimal modification, our method fits all distributions simultaneously without extra training or inference overhead. To be specific, we make the following modifications to Jen-1 (Li et al., 2023): (a) We expand the input-output architecture to encompass latent representations for multiple music tracks. This expansion enables the model to capture relationships between these tracks. (b) We introduce timestep vectors to govern the generation of each individual track. This inclusion provides flexibility for conditional generation, allowing for fine-grained control. (c) Special prompt tokens have been added to indicate specific generation tasks, reducing ambiguity and enhancing the model's performance. In addition, we propose a curriculum training strategy to progressively train the model on increasingly challenging tasks. This training regimen begins with generating a single track, then advances to handling multiple tracks, and ultimately culminates in the generation of diverse combinations of multiple music tracks.

On the other hand, current models lack the flexibility necessary for users to easily incorporate their artistic preferences into the music generation process. We contend that a more seamless integration of human creativity and AI capabilities can enhance music composition. To accomplish this, we propose the implementation of a Human-AI co-composition workflow during the model's inference phase. As illustrated in Figure 1, producers and artists collaboratively curate and blend AI-generated tracks to realize their creative visions. More specifically, our model enables the generation of tracks based on both textual prompts and satisfactory audio segments from previous iterations. Through selective re-generation guided by feedback, users can engage in an iterative collaboration with the AI until all tracks meet their desired standards. This approach complements individual artistic imagination with the generative power of AI, offering precise control tailored to individual preferences. Our evaluations demonstrate that JEN-1 Composer excels in generating a wide range of track combinations with state-of-the-art quality and flexibility.

To summarize, the contributions of this work are four-fold:

1. For the first time, we introduce an innovative workflow for collaborative music generation involving both humans and AI. This workflow is designed for the iterative creation of multi-track music.

2. We present JEN-1 Composer, a unified framework that effectively models marginal, conditional, and joint probability distributions for generating multi-track music.

3. We design an intuitive curriculum training strategy to enhance the model capacity by progressively reducing the required conditioning music information.

4. Through quantitative and qualitative assessments, we demonstrate that JEN-1 Composer achieves state-of-the-art quality and alignment in generating conditional multi-track music.

2 RELATED WORK

In this section, we introduce the research background related to this work from two aspects. First, we discuss the technical advances in conditional music generation. Then, we review key works in the field of multi-track music generation.
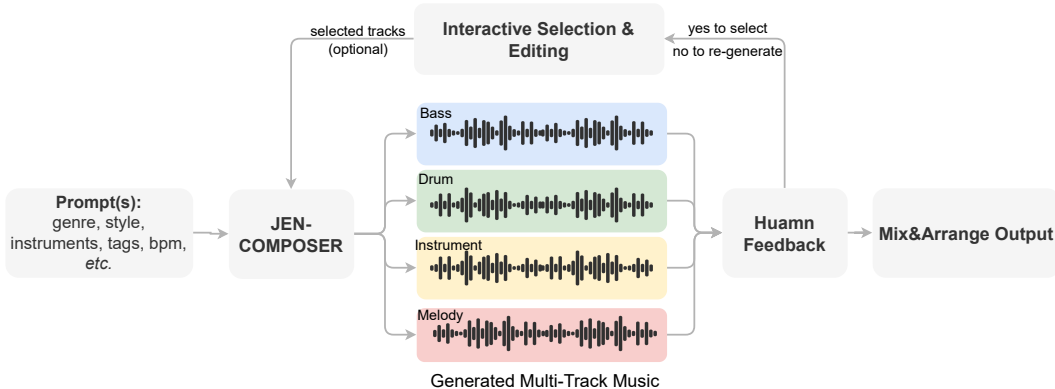
Figure 1: The Human-AI co-composition workflow of JEN-1 Composer. JEN-1 Composer generates multiple music tracks conditioned on two forms of human feedback: 1) text prompts indicating desired genres, eras, rhythms *etc.*, and 2) iterative selection/editing of satisfactory track subsets from previous generations. The selected subsets can serve as conditional signals to guide JEN-1 Composer in generating remaining tracks, ensuring contextual consistency between different tracks. This collaborative loop of human curation and AI generation is repeated until all tracks are deemed satisfactory. Finally, the tracks are mixed into a complete cohesive musical piece.

2.1 CONDITIONAL MUSIC GENERATION

In the field of music generation, conditional models have been widely applied for various tasks. Based on the temporal alignment of the conditioning signals, they can be categorized into two types: One uses low-level control signals that are highly aligned with the audio output, such as lyrics (Yu et al., 2021) and MIDI sequences (Muhamed et al., 2021). The other leverages high-level semantic representations like text (Kreuk et al., 2022; Agostinelli et al., 2023; Liu et al., 2023) or images (Huang et al., 2023b) to provide overall coherence without tight alignment. Considering the scarcity of aligned data, many models adopt self-supervised training (Marafioti et al., 2019; Borsos et al., 2023) to improve generalization. Another challenge is the high complexity of raw waveforms (Gârbacea et al., 2019), making direct generation intractable. Thus, various feature extraction and representation techniques have been extensively studied, such as VQ-VAE/GAN-based methods using mel-spectrograms (Van Den Oord et al., 2017; Creswell et al., 2018; Huang et al., 2023a), and quantization-based methods that convert waveforms into more compact discrete representations (Zeghidour et al., 2021; Défossez et al., 2022). Recently, non-autoregressive generation based on diffusion models (Ho et al., 2020) has emerged as a promising approach, where MeLoDy (Lam et al., 2023) and Jen-1 (Li et al., 2023) achieve high-fidelity music generation. Our proposed JEN-1 Composer follows this line of work, but differs from existing conditional music generation models in two aspects. In addition to textual prompts that provide high-level control over global styles, our model utilizes cross-track dependencies as an additional type of tight alignment conditioning. Specifically, we generate missing music tracks conditioned on any given combination of existing tracks, exploiting their temporal alignments to improve harmony and mixing. Furthermore, instead of directly producing a full mix, our generation target is separated music tracks, aligned with real-world music production workflows for track-wise editing and creation. The vectorized timesteps and prompt prefixes guide the model to generate each track conditioned on others, facilitating iterative refinement and Human-AI collaborative creation. In summary, our model explores multi-level conditioning signals and targets separated tracks for enhanced coherence, controllability and creativity.

2.2 MULTI-TRACK MUSIC GENERATION

Multi-track music generation is an emerging research direction. Pioneering works have explored multi-track symbolic music generation with various generative models. MuseGAN (Dong et al., 2018) presented one of the first attempts using GANs, though the results were limited in diversity. Subsequent works explore more powerful generator architectures. For instance, MIDI-Sandwich2 (Liang et al., 2019) adopted hierarchical RNNs and VAEs to model long-term track dependencies. MMM (Ens & Pasquier, 2020) and MTMG (Jin et al., 2020) incorporated the trans-

former's attention mechanisms. More recent approaches like MTT-GAN (Jin et al., 2022) combined GANs and transformers to produce multi-track music that conforms to the music rules. Unlike these approaches, we propose a diffusion model for multi-track generation. Compared to VAEs, GANs and transformers, diffusion models have shown superior performance on generative modeling of music (Kong et al., 2020; Liu et al., 2023). Moreover, we use source separation tools like Spleeter (Hennequin et al., 2020) and Demucs (Défossez, 2021; Rouard et al., 2023) to massively augment our waveform training data instead of relying on professional MIDI annotations, greatly reducing the data acquisition difficulty. In addition, we do not explicitly incorporate music theory modeling. This provides more creative freedom without confining the results to a single style. We believe this simpler framework can empower users to compose multi-track music in a more intuitive and unrestrained way. Our model JEN-1 Composer learns the inter-track dependencies and relationships in an implicit manner from the training data. By operating directly on raw audio and empowering unrestrained generation, JEN-1 Composer explores the possibility of multi-track music creation beyond the boundaries set by existing MIDI-based methods and music rules.

3 PRELIMINARY

3.1 DIFFUSION MODEL

Diffusion models (Ho et al., 2020) are a type of generative model that can generate high-quality samples via iterative denoising. A noise prediction model parameterized by θ takes the timestep t and the corrupted sample $\mathbf{x}_t$ as input. It is trained to estimate the conditional expectation $\mathbb{E}\left[\epsilon_t | \mathbf{x}_t\right]$ by minimizing the following regression loss:

$$\min_\theta \mathbb{E}_{t,\mathbf{x}_0,\epsilon_t} \left\| \epsilon_t - \epsilon_\theta\left(\mathbf{x}_t, t\right) \right\|_2^2 \tag{1}$$

where t is uniformly sampled from $\{1, 2, \ldots, T\}$ and ϵ_t is the injected standard Gaussian noise that perturbs the original data $\mathbf{x}_0$ as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t \tag{2}$$

Here, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, $\alpha_t = 1 - \beta_t$, and β_t denotes the noise schedule controlling the noise levels over time. With an optimized noise predictor, we can reversely approximate $\mathbf{x}_0$ by sampling from a Gaussian model $p\left(\mathbf{x}_{t-1}|\mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}|\boldsymbol{\mu}_t\left(\mathbf{x}_t\right), \sigma_t^2 \boldsymbol{I}\right)$ in a stepwise manner (Bao et al., 2023), where the optimal mean under maximal likelihood estimation is:

$$\boldsymbol{\mu}_t^*\left(\mathbf{x}_t\right) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta\left(\mathbf{x}_t, t\right) \right) \tag{3}$$

3.2 AUDIO LATENT REPRESENTATION

Directly modeling raw audio waveforms is intractable due to high dimensionality, where $\mathbf{x} \in \mathbb{R}^{c \times s}$ represents the waveform with c channels and s being the sequence length. To obtain a more compact representation, we first encode $\mathbf{x}$ into the latent space $\mathbf{z} \in \mathbb{R}^{d \times \hat{s}}$ using a pretrained autoencoder, where $\hat{s} \ll s$ is the compressed sequence length and d is the latent dimension:

$$\mathbf{z} = f_\phi(\mathbf{x}), \quad \widehat{\mathbf{x}} = g_\psi(\mathbf{z}) \tag{4}$$

Here f_ϕ and g_ψ denote the encoder and decoder networks respectively. By compressing the original high-dimensional waveform $\mathbf{x}$ into the lower-dimensional latent variable $\mathbf{z}$, we obtain a more compact and tractable representation for subsequent processing. In this work, we pretrain our own autoencoder model for audio reconstruction, following the Jen-1 architecture proposed in Li et al. (2023). While other external pre-trained models like SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2022) could also be compatible, we do not test them in this paper. The diffusion process operates on the latent space.

4 METHOD

In this section, we introduce the proposed methodology of JEN-1 Composer for flexible multi-track music generation. We first describe the key modifications to the Jen-1 model architecture in Section 4.1. This is followed by the curriculum training strategy in Section 4.3 and the interactive inference approach in Section 4.4.
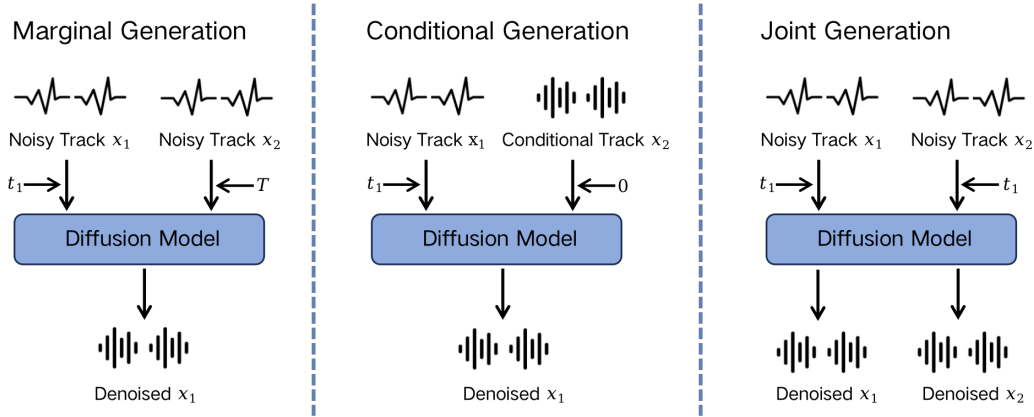
Figure 2: Illustration depicting the three distinct modes employed by JEN-1 Composer to generate one of the multi-track variables $\mathbf{x}_1$. These modes encompass marginal, conditional, and joint generation. By introducing distinct noise perturbations into their respective input tracks, our JEN-1 Composer learns the art of reconstructing and generating clean tracks across diverse settings. Specifically, for marginal generation, we introduce noise with a maximum timestep of T into all other tracks. In contrast, for conditional generation, we maintain a noise-adding level of 0 and utilize the original or generated audio from all other tracks as the condition. In the case of joint generation, all noise is independently sampled. Note that here we only present a scenario involving two tracks in this illustration for simplicity.

4.1 MULTI-TRACK MUSIC GENERATION

To enable JEN-1 Composer to handle multi-track input and output for joint modeling, we make minimal extensions to its original single-track architecture. As elaborated below, the input-output representation, timestep vectors, and prompt prefixes are adapted to fit multi-track distributions efficiently using a single model.

4.1.1 MULTI-TRACK INPUT-OUTPUT REPRESENTATION

We extend the single-track input $\mathbf{x} \in \mathbb{R}^{c \times s}$ of Jen-1 to multi-track inputs $\boldsymbol{X} = \left[\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k \right]$, where $\mathbf{x}^i \in \mathbb{R}^{c \times s}$ is the waveform for the i-th track and k is the total number of tracks. The waveform of each track $\mathbf{x}^i$ is encoded into the latent space using the pretrained encoder f_ϕ, namely $\mathbf{z}^i = f_\phi(\mathbf{x}^i) \in \mathbb{R}^{d \times \hat{s}}$. The input tracks are concatenated along the channel dimension to form the final input $\boldsymbol{Z} \in \mathbb{R}^{kd \times \hat{s}}$. Correspondingly, the single-track output in Jen-1 is expanded to kd channels, then producing separate waveforms for k tracks. Extending the input-output representation to multi-track allows explicitly modeling the inter-dependencies and consistency between different tracks, which is essential for high-quality multi-track generation but lacking in single-track models. The concatenated latent representations align the structure with the multi-track waveform outputs, enabling synchronized generation across tracks. Modeling relationships among tracks also facilitates generating certain tracks conditioned on others, a key capability in flexible music creation workflows.

4.1.2 INDIVIDUAL TIMESTEP VECTORS

Along with the expanded input-output structure, we introduce separate timesteps for each track to gain fine-grained control over the generation process. To be specific, the scalar timestep t in the original Jen-1 is extended to a multi-dimensional timestep vector $[t_1, \ldots, t_k]$, where each element $t_i \in \{0, 1, \ldots, T\}$ corresponds to the noise level for the i-th track. In particular, $t_i = 0$ indicates the i-th track is given as conditional input without noise. $t_i > 0$ means the corresponding track needs to be generated by the model based on the conditional tracks. $t_i = T$ represents the maximum noise level that cannot provide any conditioning signal. As shown in Figure 2, by controlling the timestep vectors, our model can flexibly specify the tracks to reconstruct or generate for a given input, avoiding the need to retrain models for every combination of conditional tracks. This greatly improves the flexibility and reduces the training overhead. Varying timesteps for different tracks

also allows controlling the noise levels independently, making the model adaptive to more diverse generation tasks.

4.2 Integrating Task Tokens as Prefix Prompts

In addition to the conventional text prompts describing the music content and style, we incorporate task-specific tokens as prefixes to guide the generation process. These task tokens serve as explicit directives for the model, offering clear instructions regarding the current generation task, akin to the use of text prompts for controlling musical style. By utilizing these task-specific prefixes, we enhance the model's capability to focus its generative efforts on producing content that aligns with the specified task, thus reducing ambiguity and elevating the quality of output. To illustrate this concept, consider the utilization of prompt prefixes such as "[bass & drum generation]". These prefixes effectively communicate to the model the immediate generation objective, in this case, the generation of bass and drum tracks. This explicit task signaling enables the model to concentrate its generative capacity on crafting these missing tracks while taking into account the existing conditional tracks. Through the integration of task-specific prefixes, accompanied by enhanced individual timestep vectors, our proposed JEN-1 Composer demonstrates a remarkable capacity to efficiently model the marginal, conditional, and joint probability distributions associated with the various tracks. All these tasks are addressed within a single, unified model, a testament to the versatility and adaptability of our approach in handling multifaceted generative challenges.

4.3 Progressive Curriculum Training Strategy

We propose a curriculum training strategy to progressively enhance the model's capability in modeling joint and conditional distributions over k tracks. The strategy starts by reconstructing audio with only one missing track. It then steadily increases the number of tracks to be generated in each training step, thus enhancing the difficulty. Critically, instead of completely replacing easier stages, we gradually increase the probability that more challenging stages are selected during training. All stages, representing tasks with varying difficulties, are trained with designated probabilities. In this manner, the model is steadily presented with more difficult modeling tasks, while continually being trained on simpler tasks to avoid forgetting.

The schedule consists of k stages:

- Stage 1: Reconstruct 1 random track out of k per step, with $k-1$ tracks given as conditional inputs.
- Stage 2: Generate 2 random tracks out of k per step, conditioned on the other $k-2$ tracks.
- ...
- Stage k: Free generation of all k tracks without any conditional tracks.

This curriculum not only ensures the model learns basic reconstruction skills but also gently enhances its capacity in coordinating more tracks simultaneously. By incrementally growing the task difficulty, it prevents the model from overfitting simple cases while forgetting more complex generation behaviors, a common issue in conventional training. The progressive schedule allows smooth transitioning of the model from reconstructing existing combinations to freely imagining novel mixtures of tracks.

4.4 Interactive Human-AI Co-composition Workflow

During inference, our model supports conditional generation of multiple tracks given 0 to $k-1$ tracks as input conditions. To enable Human-AI collaborative music creation, we devise an interactive generation procedure, detailed in Algorithm 1.

The proposed interactive inference approach seamlessly combines human creativity with AI capabilities to enable collaborative music generation. During the iterative process, humans can focus on improvising particular tracks that pique their interest, while maintaining harmony and consistency with the overall generation guided by the model. This complementary Human-AI workflow is aligned with real-world music composition practices, and provides the following benefits:

Algorithm 1 Interactive Human-AI Co-composition Workflow

User provides a text prompt $\mathbf{p}$
Model generates k-track audio $(\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2, \ldots, \hat{\mathbf{x}}^k)$ conditioned on $\mathbf{p}$
User selects satisfactory track subset $\mathbb{S}$
repeat
    Model generates tracks conditioned on $\mathbb{S}$ and $\mathbf{p}$
    User selects satisfactory tracks from generation and adds them to $\mathbb{S}$
until all k tracks selected

- It allows progressively layering and polishing each track with a closed-loop human feedback mechanism, facilitating nuanced refinement difficult for pure AI generation.

- With humans picking satisfactory samples at each iteration, it helps filter out low-quality samples and steer the generation towards desirable directions.

- By interacting with human creators and incorporating their inputs, the model can keep improving its understanding of human aesthetic preferences and sound quality standards.

- The generation can leverage both human ingenuity and AI capabilities. Humans excel at creative improvisation while AI provides helpful cues to ensure coherence and prompt-consistency.

- The collaborative experience enhances engagement and sense of control for human producers. It enables realizing their creative visions through an AI assistant.

In summary, the interactive inference paradigm organically couples human creativity with AI generation to enable enhanced music co-creation. It balances open-ended improvisation and overall structural coherence, combining the strengths of both to take music generation to the next level.

5 EXPERIMENT

5.1 SETUP

Datasets. We employ a private studio recording dataset containing 800 hours of high-quality multi-track audio data to train JEN-1 Composer. The dataset consists of 5 types of audio tracks that are temporally aligned, including bass, drums, instrument, melody, and the final mixed composition. All tracks are annotated with unified metadata tags describing the genre, instruments, moods/themes, tempo, *etc.* To construct the training and test sets, we first randomly split the dataset into a 4:1 ratio. We then extract aligned segment snippets from the 5 tracks using the same start and end times to preserve temporal consistency. This process ensures the multi-track snippets in our dataset are temporally synchronized for training the model to learn cross-track dependencies and consistency. The training set encompasses 640 hours of audio data, spanning a diverse array of musical styles and instrumentation. In contrast, the remaining test set comprises 160 hours of audio, serving as the basis for evaluating the model's ability to generalize. With the presence of comprehensive annotations and temporal alignment, our dataset plays a pivotal role in training JEN-1 Composer. It equips the model with the capability to generate high-quality multi-track music in response to text prompts that convey desired attributes.

Evaluation Metrics. We have conducted a comprehensive evaluation of our methodology, encompassing both quantitative and qualitative dimensions. For quantitative metrics, we adopt the CLAP score (Elizalde et al., 2023) to measure the alignment between text and music tracks. More specifically, we have computed CLAP scores for both the mixed track and each individual separated track. In the case of JEN-1 Composer, we have simply summed the four generated tracks to derive the mixed track and subsequently computed the Mixed CLAP score. For state-of-the-art models that directly generate mixed audio, we adopt Demucs (Défossez, 2021; Rouard et al., 2023) to separate the mixed tracks prior to calculating per-track CLAP scores. For qualitative analysis, we employ a Relative Preference Ratio (RPR) from human evaluation to assess the quality of mixed audio generated by different models. Specifically, we generated samples from various models in response to text prompts, and had multiple human raters compare these sample pairs, recording the percentage of times a model's generation was preferred over JEN-1 Composer's mix. A higher RPR (ranging from

Table 1: Multi-track text-to-music generation. We compare objective and subjective metrics for JEN-1 Composer against a number of state-of-the-art baselines. We utilize the open-source model whenever feasible, and for MusicLM, we rely on the publicly accessible API.

METHODS	CLAP↑					RPR↑
	BASS	DRUMS	INSTRUMENT	MELODY	MIXED	MIXED
MusicLM	0.16	0.17	0.23	0.28	0.28	27%
MusicGen	0.17	0.15	0.25	0.33	0.35	36%
JEN-1	0.19	0.16	0.29	0.32	0.36	40%
JEN-1 Composer	**0.21**	**0.18**	**0.29**	**0.36**	**0.39**	–

0% to 100%) indicates a stronger preference for a given model over JEN-1 Composer's mix. Our evaluation process emphasized aspects including coherence, logical consistency, and smoothness of quality across the generated tracks.

Implementation Details. Our multi-track music generation task encompasses four distinct tracks: bass, drums, instrument, and melody, as well as the composite mixed track. All audio data are high-fidelity stereo audio sampled at a rate of 48 kHz. Specifically, we employ a hop size of 320 to encode the audio, resulting in a latent space representation of 150 frames per second, each comprising 128 dimensions. The intermediate dimension within the cross-attention layers is configured to be 1024. Prior to compression into the latent space, we adjust the volumes of individual tracks by scaling them in accordance with the mixing volumes, ensuring that their relative loudness remains consistent. Semantic understanding of the text prompts is achieved through the utilization of the pre-trained FLAN-T5 model (Chung et al., 2022).

Regarding model architecture, we make minimal modifications to Jen-1 (Li et al., 2023). As described in Section 4.1, the primary changes pertain to the input-output handling, where we concatenate the four tracks in a channel-wise manner. These tracks collectively share a 1D UNet backbone (Ronneberger et al., 2015). The single-track timestep is expanded into a timestep vector, allowing the addition of varied noise levels to each track. In the training process, for each batch, we first uniformly sample one of the four tracks at random, then assign it a non-zero timestep t_i, sampled from $\{1, \ldots, T-1\}$, which determines the strength of Gaussian noise injected into the track's latent embedding. The timesteps and noise levels for the other three tracks are stochastically drawn from $\{0, t_i, T\}$. Specifically, a timestep of 0 represents a clean track, which serves as the conditional signal for guided generation. A timestep of T signifies maximum noise level, so this track does not provide conditional guidance and hence supports unconstrained generation from the marginal distribution. Lastly, a timestep of t_i indicates that this track is jointly optimized as one of the generation targets together with the selected i-th track. This unified framework comprehensively covers all permutations of multi-track generation tasks. Additionally, we employ classifier-free guidance (Ho & Salimans, 2022) to enhance the correlation between generated tracks and text prompts. JEN-1 Composer is trained on two A100 GPUs, with other hyperparameters including AdamW optimizer (Loshchilov & Hutter, 2017), a linear decay learning rate initialized at $3e^{-5}$, batch size of 12, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay of 0.1, and gradient clipping threshold of 0.7.

5.2 COMPARISON WITH STATE-OF-THE-ARTS

To the best of our knowledge, our proposed JEN-1 Composer represents the first attempt to address the challenging task of authentic multi-track music generation. In this context, we undertake a comparative examination with other state-of-the-art text-to-music generation approaches, namely MusicLM (Agostinelli et al., 2023), MusicGen (Copet et al., 2023), and Jen-1 (Li et al., 2023). It is worth noting that all these methods are limited to generating single-track music with mixed attributes. As demonstrated in Table 1, JEN-1 Composer achieves superior performance over other state-of-the-art methods. Benefiting from its track-wise generation and flexible conditional modeling capabilities, JEN-1 Composer obtains consistently higher CLAP scores on each individual track, indicating stronger fine-grained control and alignment during multi-track generation. Consequently, the overall mixing and composition quality of JEN-1 Composer is markedly higher based on both human evaluation and quantitative metrics. Specifically, JEN-1 Composer outperforms other mod-

Table 2: Ablation studies evaluation. Starting from the baseline, we incrementally modify the configuration to investigate the effect of each component.

	CLAP↑					RPR↑
METHODS	BASS	DRUMS	INSTRUMENT	MELODY	MIXED	MIXED
baseline	0.20	0.18	0.20	0.28	0.28	16%
+ individual timestep vector	0.19	0.18	0.22	0.32	0.33	20%
+ curriculum training strategy	0.21	0.17	0.26	0.35	0.37	35%
+ interactive inference	**0.21**	**0.18**	**0.29**	**0.36**	**0.39**	–

els by a substantial margin in the CLAP score of the mixed track, demonstrating its advantage in coherently coordinating different tracks guided by the text prompts. Meanwhile, the results on the RPR metric also show users' strong preference towards mixes generated by JEN-1 Composer compared to other models. In summary, conditional multi-track generation enables JEN-1 Composer to achieve state-of-the-art performance and generate satisfying music aligned with the textual descriptions. The unified modeling approach provides an elegant solution for controlling inter-track relationships.

5.3 ABLATION STUDIES

We have conducted ablation studies to ascertain the effectiveness of key components within JEN-1 Composer. The findings, detailed in Table 2, originate from an initial vanilla baseline model featuring a four-track input/output structure inspired by Jen-1. We then progressively add the proposed techniques row by row. First, using individual timestep vectors for each track is crucial for modeling marginal and conditional distributions, instead of only joint distribution in the baseline. This leads to substantially higher CLAP scores on individual tracks. Second, the curriculum training strategy facilitates a smooth transition from learning simple conditional models to complex joint generation, further improving results, especially on challenging tracks like melody and instrument. Finally, interactively combining with the Human-AI co-composition workflow yields the best mixing quality, as the model can flexibly switch between modes with multiple injections of human preference. The extra conditional signals from feedback guide the model to overcome weaknesses and generate high-quality results for all tracks. For example, it can first generate drums and bass, then leverage the conditional distribution to produce satisfactory melody and instrument conditioned on them. In summary, benefiting from the dedicated design, JEN-1 Composer boasts flexibility in fine-grained conditional control and achieves promising generation quality for multi-track music synthesis.

6 CONCLUSION

In this study, we introduce JEN-1 Composer, a comprehensive framework for multi-track music generation that harnesses the capabilities of diffusion models. This framework extends the single-track architecture of Jen-1, enabling efficient handling of marginal, joint, and conditional distributions across multiple tracks within a unified model. Moreover, we propose a curriculum training strategy designed to promote stable training, progressing from basic reconstruction to unconstrained composition. Notably, our work also presents a novel interactive Human-AI co-composition workflow. Comprehensive evaluations, including quantitative metrics and human assessments, demonstrate its exceptional performance in high-fidelity music generation while offering versatile control over the creative process.

Although our generative modeling of JEN-1 Composer has made significant advances, limitations remain, particularly in its ability to produce audio that meets specific aesthetic and music theory directives compared to professional music production. Truly realizing AI-aided music creativity necessitates deeper collaboration between engineering, design, and art to create intuitive Human-AI co-creation interfaces and experiences. Moving forward, we are enthusiastic about exploring this landscape and jointly developing innovative techniques and workflows to unlock the creative potential of human-machine partnerships. By enhancing the connections between technology and artistry, we envision AI as an inspiring collaborator for limitless musical creativity.

REFERENCES

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

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv preprint arXiv:2303.06555*, 2023.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35 (1):53–65, 2018.

Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

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

Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pp. 1068–1077. PMLR, 2017.

Jeff Ens and Philippe Pasquier. Mmm: Exploring conditional multi-track music generation with the transformer. *arXiv preprint arXiv:2008.06048*, 2020.

Emma Frid, Celso Gomes, and Zeyu Jin. Music creation by example. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–13, 2020.

Cristina Gârbacea, Aäron van den Oord, Yazhe Li, Felicia SC Lim, Alejandro Luebs, Oriol Vinyals, and Thomas C Walters. Low bit-rate speech coding with vq-vae and a wavenet decoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 735–739. IEEE, 2019.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5 (50):2154, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018.

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

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023b.

Cong Jin, Tao Wang, Shouxun Liu, Yun Tie, Jianguang Li, Xiaobing Li, and Simon Lui. A transformer-based model for multi-track music generation. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 11(3):36–54, 2020.

Cong Jin, Tao Wang, Xiaobing Li, Chu Jie Jiessie Tie, Yun Tie, Shan Liu, Ming Yan, Yongzhi Li, Junxian Wang, and Shenze Huang. A transformer generative adversarial network for multi-track music generation. *CAAI Transactions on Intelligence Technology*, 7(3):369–380, 2022.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

Max WY Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, et al. Efficient neural music generation. *arXiv preprint arXiv:2305.15719*, 2023.

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

Xia Liang, Junmin Wu, and Jing Cao. Midi-sandwich2: Rnn-based hierarchical multi-modal fusion generation vae networks for multi-track symbolic music generation. *arXiv preprint arXiv:1909.03522*, 2019.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Andrés Marafioti, Nathanaël Perraudin, Nicki Holighaus, and Piotr Majdak. A context encoder for audio inpainting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12): 2362–2372, 2019.

Aashiq Muhamed, Liang Li, Xingjian Shi, Suri Yaddanapudi, Wayne Chi, Dylan Jackson, Rahul Suresh, Zachary C Lipton, and Alex J Smola. Symbolic music generation with transformer-gans. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 408–417, 2021.

Christine Payne. Musenet, 2019. *URL https://openai.com/blog/musenet*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention– MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 23*, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Yi Yu, Abhishek Srivastava, and Simon Canales. Conditional lstm-gan for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1):1–20, 2021.

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

Hongyuan Zhu, Qi Liu, Nicholas Jing Yuan, Kun Zhang, Guang Zhou, and Enhong Chen. Pop music generation: From melody to multi-style arrangement. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(5):1–31, 2020.