JEN-1 DreamStyler: Customized Musical Concept Learning via Pivotal Parameters Tuning

Boyu Chen, Peike Li*, Yao Yao, Alex Wang

Jen Music AI
boyu@jenmusic.ai, alex@jenmusic.ai

Abstract

Large models for text-to-music generation have achieved significant progress, facilitating the creation of high-quality and varied musical compositions from provided text prompts. However, input text prompts may not precisely capture user requirements, particularly when the objective is to generate music that embodies a specific concept derived from a designated reference collection. In this paper, we propose a novel method for customized text-to-music generation, which can capture the concept from a two-minute reference music and generate a new piece of music conforming to the concept. We achieve this by fine-tuning a pretrained text-to-music model using the reference music. However, directly fine-tuning all parameters leads to overfitting issues. To address this problem, we propose a Pivotal Parameters Tuning method that enables the model to assimilate the new concept while preserving its original generative capabilities. Additionally, we identify a potential concept conflict when introducing multiple concepts into the pretrained model. We present a concept enhancement strategy to distinguish multiple concepts, enabling the fine-tuned model to generate music incorporating either individual or multiple concepts simultaneously. We also introduce a new dataset and evaluation protocol for this task. Our proposed JEN1-DreamStyler outperforms several baselines in both qualitative and quantitative evaluations.

Demo — https://www.jenmusic.ai/audio-demos

Datasets —
https://huggingface.co/datasets/JenMusicAI/DreamStyler

Introduction

Recent advancements in generative models (Vaswani et al. 2017; Nichol and Dhariwal 2021; Rombach et al. 2022) have marked significant progress in the field of text-to-music generation (Agostinelli et al. 2023; Copet et al. 2023; Liu et al. 2023; Li et al. 2024; Yao et al. 2023; Schneider et al. 2023; Liu et al. 2024; Chen et al. 2024; Evans et al. 2024a,b; Ziv et al. 2024). These models, usually trained on large-scale datasets of text-music pairs, allow users to experience a novel form of musical interaction, where they can input a textual description and receive a piece of music that aligns with the described mood, genre, theme, *etc.* The vastness and diversity of the training datasets enable these models to handle a wide range of musical concepts, including contents (*e.g., instruments*) and styles (*e.g., genres*).

Despite their comprehensive training, text-to-music generation models (Li et al. 2024) face significant challenges in fully capturing and replicating the broad spectrum of human musical concepts, which often exhibit a long-tailed distribution (Celma Herrada et al. 2009). Specifically, the models struggle with unique or context-specific musical concepts that appear infrequently and may not be included in their training datasets. For example, complex melodies produced by street performers using unconventional instruments, such as water cups, or the unique timbre of a ventriloquist performing alone, frequently lack accurate textual descriptions. In light of these limitations, the pursuit of customized music generation becomes increasingly significant. This highlights the vast potential and yet-to-be-realized capabilities of current text-to-music technologies.

In this work, we concentrate on customized text-to-music diffusion models by adapting them to interpret and reproduce new musical concepts, as shown in Fig. 1. Specifically, we aim to modify an existing model to accurately recognize and reproduce a particular musical concept, such as an instrument or genre, using only two minutes of reference music without textual description of the desired concept. For this purpose, leveraging the pretrained text-to-music models for direct fine-tuning offers a straightforward approach. Nevertheless, this method faces two significant challenges. Firstly, there is a tendency for the model to overfit the given reference music, resulting in generated music that lacks diversity and closely resembles the reference. Secondly, directly fine-tuning the model to incorporate multiple musical concepts simultaneously proves to be impractical. For instance, when attempting to merge distinct sounds from a piano and a guitar from two separate reference tracks, the model often suffers from a concept conflict issue, where one concept dominating the generation, ignoring the other, which impedes the model's capacity to effectively combine multiple musical concepts coherently.

To tackle these challenges, we propose the JEN-1 DreamStyler, introducing an innovative regularization method named Pivotal Parameters Tuning. This method selectively fine-tunes concept-specific pivotal parameters within the

*Work done while was at Jen Music AI

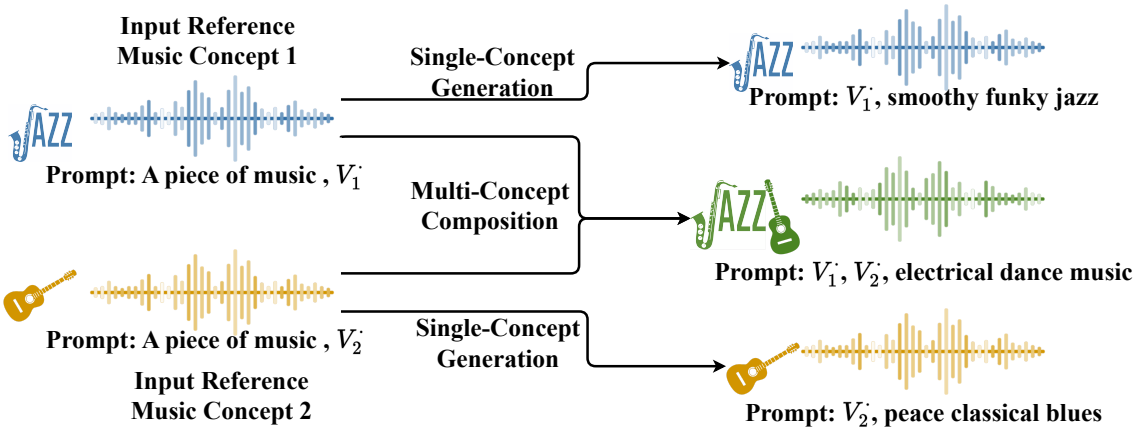


Figure 1: Utilizing a mere two minutes of reference music representing a new concept, our proposed JEN-1 DreamStyler can understand and reproduce the musical concept. Reference musical concepts could be an instrument (*e.g. guitar*), a genre (*e.g.*, jazz), *etc.* Our JEN-1 DreamStyler works on mastering a single musical concept as well as multiple musical concepts.

network, maintaining the remainder unchanged. It employs a sparse mask to identify the most pivotal parameters, based on their variation relative to the reference music. The underlying principle asserts that parameters exhibiting greater variation in mask values are pivotal for generating the target musical concept. Consequently, these pivotal parameters are selected for subsequent fine-tuning, while the remaining parameters are kept non-trainable. By adopting this strategy, our model effectively learns the musical concept from the reference and preserve generality of the pretrained model.

Beyond selectively tuning network parameters, our JEN-1 DreamStyler incorporates trainable identifier tokens into the input prompt. The conventional text-inversion method (Gal et al. 2022) utilizes a single extra token for each musical concept, *e.g.*, 'A piece of V* music'. However, this method is inadequate when dealing with multiple concepts, *e.g.*, 'A piece of V_1^* and V_2^* music'. We observed that distinct tokens V_1^* and V_2^*, despite their initial uniqueness, eventually converge into highly similar after processing by the text encoder. To resolve this issue, our model innovates by assigning multiple tokens to each musical concept. This strategy significantly diversifies the representation of each concept, ensuring tokens corresponding to different concepts remain distinct even after the text encoder.

To validate our method, we introduce a new benchmark dataset and an evaluation protocol. Through a combination of qualitative and quantitative assessments, we demonstrate the effectiveness of our proposed method. To summarize, the contributions of this work are multi-dimensional:

- We introduce an innovative framework designed specifically for data-efficient, customized music generation. With the two-minute reference music, the framework can generate diverse music with unique musical concept.
- Our approach incorporates Pivotal Parameters Tuning method, which selectively tuning the pivotal parameters for generating the specific musical concept, effectively addresses the challenge of over-fitting.
- We tackle the challenge of concept conflict during introducing multiple concepts. The concept enhancement strategy ensures each concept is distinctly and effectively represented within the text-to-music generation model.
- To support this challenging task, we develop a novel dataset and evaluation protocol specifically tailored for customized music generation. This dataset serves as a benchmark for assessing our method and establishes a foundation for future research in this area.

Related Work

Text-to-Music Generation. Text-to-music generation focuses on converting textual descriptions into corresponding musical compositions. This interdisciplinary area merges language descriptions with musical creativity, leveraging generative models to produce music that reflects themes, moods, or tags expressed in text. Recent advancements in text-to-music generation have shown promising results. Riffusion (Forsgren and Martiros 2022), for instance, has adapted Stable Diffusion model for music generation. By converting music into mel-spectrograms, Riffusion transforms challenging text-to-music generation into a more manageable text-to-image task. MusicGen (Copet et al. 2023) utilizes a transformer-based autoregressive model, producing music through discrete tokens. Its innovative delay pattern technique significantly boosts the efficiency of music generation. Furthermore, JEN-1 (Li et al. 2024) proposes a multi-task training framework, based on a diffusion model, that uniquely combines autoregressive and non-autoregressive training. This integration results in the production of high-fidelity stereo music, demonstrating versatility and advancement in this field. Despite these technological advancements, text-to-music generation still faces substantial challenges, such as the difficulty in formulating accurate and detailed text descriptions that align with user preferences. To address this, we proposes a customized music generation method which is capable of generating vari-

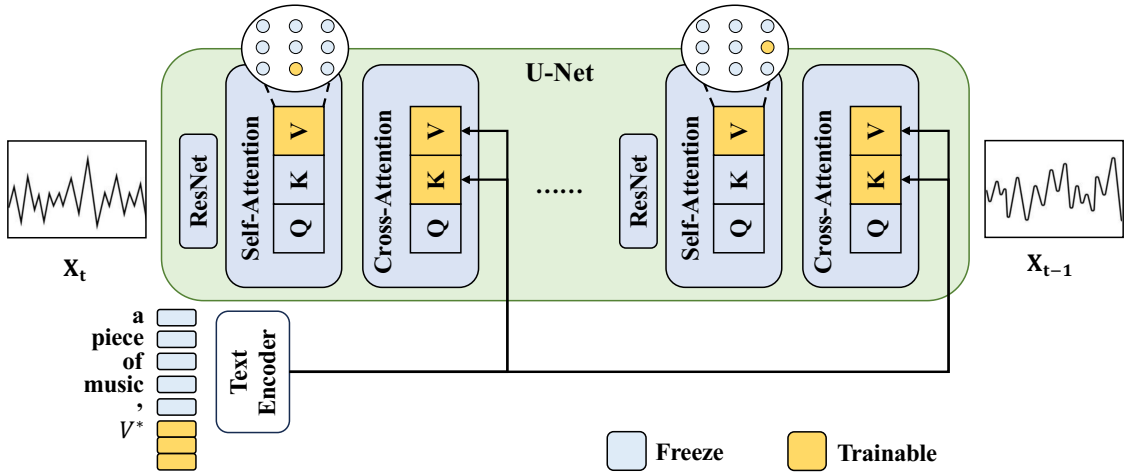


Figure 2: Given reference music, we select and fine-tune the most pivotal parameters within the U-Net module of our text-to-music diffusion model. Furthermore, we introduce several trainable concept identifier tokens, denoted as V^*, to present these new concepts. During training, we efficiently tune these pivotal value projection parameters in the self-attention layers and all key and value projection parameters in the cross-attention layers, in conjunction with the concept identifier tokens. For simplicity, we only illustrate scenarios involving the learning of a single musical concept.

ous music pieces based on reference music. This approach overcomes the challenges of text description dependency, offering a more flexible and user-friendly solution for customized music generation.

Customized Creation using Diffusion Models. Customized Creation in image generation using diffusion models has become a highly popular area of research. This approach focuses on generating images that either share a style or contain objects similar to those in reference images. Numerous works have contributed significantly to the development of this field. For instance, Text Inversion (Gal et al. 2022) has innovated by adding new pseudo-words to the vocabulary of a frozen text-to-image model. This allows the model to represent a unique concept with just a single word embedding, effectively capturing a wide range of diverse and distinct ideas. Dreambooth (Ruiz et al. 2023) further expands on this by introducing a method to associate unique identifiers with specific subjects. By training the entire U-Net (Ronneberger, Fischer, and Brox 2015) with their class-specific prior preservation loss, Dreambooth prevents the model from catastrophic forgetting and language drift (Lee, Cho, and Kiela 2019) and enables the creation of photo-realistic images of these subjects in a variety of contexts and poses. Additionally, Custom Diffusion (Kumari et al. 2023) has enhanced training efficiency by focusing on training only a portion of the parameters and utilizing regularization samples from the training dataset. They also propose a new regularization technique for multi-concept training. In the field of music generation, (Plitsis et al. 2024) investigate this task with two established methods above. Different from (Plitsis et al. 2024), we propose new techniques specifically designed for music generation, including pivotal parameter tuning and a concept enhancement strategy. Besides, we explore multiple concept learning which is not covered in (Plitsis et al. 2024). Furthermore, we introduce a new dataset and an evaluation method, thus laying the groundwork for future developments in this burgeoning field.

Preliminary

Diffusion Model

In this work, we employ the JEN-1 model (Li et al. 2024) as our foundation model, which is a state-of-the-art text-to-music generation model built upon the diffusion models. Diffusion models, such as those described by (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021), represent probabilistic generative models designed to approximate complex data distributions. These models operate by transforming simple noise distributions into intricate representations, a process particularly effective in high-quality generation.

The diffusion model is anchored in two primary processes: forward diffusion and reverse diffusion. In the forward diffusion phase, the model incrementally introduces Gaussian noise into the data over a series of steps. Each step in this Markov Chain can be mathematically expressed as

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where x_t is the data at time step t and β_t are predefined noise levels. Conversely, the reverse diffusion process involves a gradual denoising of the data. This is achieved through a neural network that learns to reverse the noise addition, a key element in synthesizing realistic audio. The reverse process can be described by the equation

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta^2(t)\mathbf{I}), \tag{2}$$

where the functions μ_θ and σ_θ^2 are parameterized by the neural network, enabling the precise prediction of mean and variance at each reverse diffusion step.

The learning mechanism of diffusion models entails a fine balance between the forward diffusion process and the reverse denoising process. The latter utilizes a noise prediction model, parameterized by θ, to estimate the conditional expectation $\mathbb{E}[\epsilon_t|x_t]$ by minimizing a regression loss. This loss, expressed as

$$\min_{\theta} \mathbb{E}_{t,x,\epsilon} \left[\|\epsilon_t - \epsilon_\theta(x_t, t)\|_2^2 \right], \tag{3}$$

where ϵ_t represents stochastic noise at timestep t, and $\epsilon_\theta(\cdot)$ denotes a time-conditional 1D U-Net, guides the model in learning distribution of original data from its noisy version.

Text-to-Music Generation

In our method, JEN-1 serves as the foundational model for text-to-music generation, which is built based on Latent Diffusion Model (LDM). This model adheres to the same forward of diffusion models mentioned, while backward process and loss function are different by incorporating textual condition y within latent space to control synthesis process,

$$\min_{\theta} \mathbb{E}_{t,x,\epsilon,y} \left[\|\epsilon_t - \epsilon_\theta(x_t, t, y)\|_2^2 \right], \tag{4}$$

where x_t is the noisy music latent input at timestep t, which is generated from the original music latent x_0.

Give the latent music feature f, the textual condition y is then integrated into the U-Net's intermediate layers via a cross-attention mechanism, defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d'}}\right) \cdot V, \tag{5}$$

where,

$$Q = W_Q \cdot f, \quad K = W_K \cdot y, \quad V = W_V \cdot y. \tag{6}$$

The matrices W_Q, W_K and W_V denote learnable projection parameters of the cross-attention layer. d' is the output dimension of key and query features. During inference, only the U-Net $\epsilon_\theta(\cdot)$ is used to synthesize the desired music generation based on the textual prompt input.

Methodology

Our proposed JEN-1 DreamStyler is designed for customized text-to-music generation. In this section, we first introduce the task, *i.e.*, customized text-to-music generation. Then, we show our proposed Pivotal Parameters Tuning to efficiently learn musical concepts. Finally, we show how to improve the quality of multiple-concept generated music.

Customized Text-to-Music Generation

We propose a customized text-to-music generation method, aiming to understand and reproduce a new musical concept from the given reference, even without any additional textual descriptions on the concept. After integrating the new concept into the pretrained text-to-music generation model, we can utilize any text prompt to generate the music with the specific concept, such as an instrument or a given genre. The generated music will be consistent with the input text prompts, as well as the learned concept. The whole pipeline of our method is shown in Fig. 2.

Fully fine-tuning. With the pretrained JEN-1 model and a two-minute music clip, the intuitive approach for concept extraction is to fine-tune JEN-1 using the music clip. However, direct fine-tuning risks overfitting to this limited dataset, leading to a loss of the generalization ability. Regularization in neural network training effectively prevents overfitting. However, the class-specific prior preservation loss, as used in (Ruiz et al. 2023) and (Kumari et al. 2023), requires object class information, which is absent in our task.

Parameter-efficient fine-tuning. To solve the overfitting problem in fully fine-tuning, prior research (Kumari et al. 2023), has demonstrated the significance of cross-attention layers during the fine-tuning process and concentrated on training only the cross-attention layers, including W_K and W_V in Eq.(6). Nevertheless, only training the cross-attention layers is insufficient for JEN-1 to effectively learn new concepts from the input reference music as the Table 1 shows.

Concept Identifier Token. To enhance concept extraction, we introduce a learnable concept identifier token, denoted as V^*, to represent the unique characteristics of the reference music. During training or generation, the concept identifier token V^* is integrated with the original textual condition y as $concat(\text{V}^*, y)$. Subsequently, this modification leads to an adaptation of the loss function. The original loss function, as defined in Eq. (4), is reformulated as follows:

$$\min_{\theta, \text{V}^*} \mathbb{E}_{t,x,\epsilon,\text{V}^*} \left[\|\epsilon_t - \epsilon_\theta(x_t, t, concat(\text{V}^*, y))\|_2^2 \right]. \tag{7}$$

Here, model parameters θ and concept identifier token V^* are trained jointly. It should be mentioned that we may utilize more than one token to represent a new concept. For simplicity, we will still use V^* to represent one concept in the following.

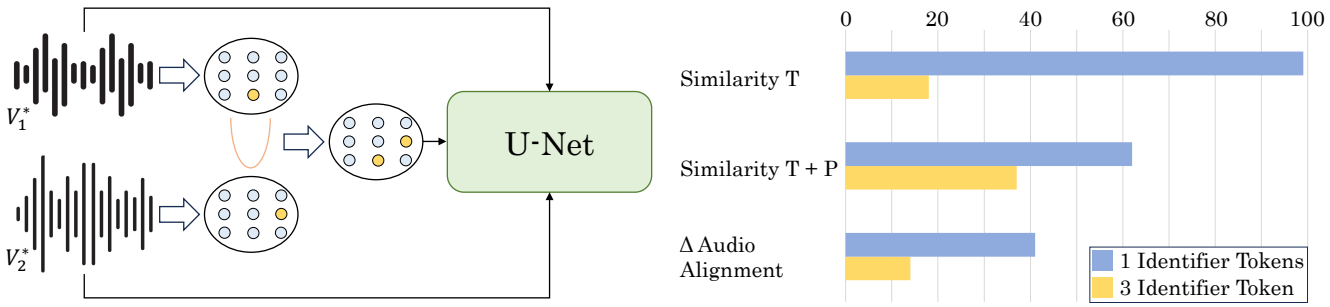
Pivotal Parameters Tuning

Training only W_K and W_V in cross-attention layers, as done in (Kumari et al. 2023), is insufficient for our model to effectively capture concepts in reference music. To enhance our model's ability to capture concepts, we introduce training W_V in self-attention layers, in addition to concept identifier tokens. However, training all W_V parameters can lead to overfitting, presenting a challenge in balancing concept capture and overfitting avoidance.

To tackle this issue, we introduce a Pivotal Parameters Tuning method, which selects the pivotal parameters of W_V in self-attention layers for optimization. We begin by initializing a trainable mask M_V, which shares the same shape as W_V in the self-attention block. This mask is subsequently element-wise multiplied with W_V, rendering the mask M_V trainable through the U-Net's forward and backward processes. All elements in the mask M_V are initialized to one, ensuring that all parameters of W_V are unchanged at the beginning. Subsequently, M_V is trained using the objective,

$$\min_{M_V} \mathbb{E}_{t,x,\epsilon,\text{V}^*} \left[\|\epsilon_t - \epsilon_{\{\theta, M_V\}}(x_t, t, concat(\text{V}^*, y))\|_2^2 \right], \tag{8}$$

where the network parameters θ and the concept identifier token V^* are fixed during training.



(a) (b)

Figure 3: (a) Framework for multiple-concept training. (b) Cosine similarity comparison of single concept identifier token and multiple concept identifier tokens from two different aspects. Besides, we show the discrepancy of two concepts Audio Alignment Score (ΔAudio Alignment), showing the distinguishing ability.

After several epochs of training the mask M_V, a refined mask M_V^* is obtained. We then compute the mask variation as $\Delta_M = |M_V - M_V^*|$. For each parameter in W_V, Δ_M represent the variation of that parameter. We select the top $P\%$ of positions with the highest values in Δ_M and designate the corresponding parameters in W_V as pivotal parameters, which will be optimized in the following training. These pivotal parameters, along with W_K and W_V from the cross-attention layers, form the trainable parameter set θ_T. The remaining parameters are treated as non-trainable parameters, denoted θ_N. The final training loss is defined as:

$$\min_{\theta_T, \mathrm{V}^*} \mathbb{E}_{t,x,\epsilon,\mathrm{V}^*} \left[\|\epsilon_t - \epsilon_{\{\theta_T,\theta_N\}}(x_t, t, concat(\mathrm{V}^*, y))\|_2^2 \right]. \tag{9}$$

Multiple Concepts Integration

Joint Training on Multiple Concepts. As Fig. 3a demonstrates, to integrate multiple concepts, we first learn the mask for each concept individually and then merge the binary masks into a new mask to determine pivotal parameters for tuning. Subsequently, we combine the training datasets for each concept and optimize pivotal parameters on the merged datasets. To distinguish each concept, we use different concept identifier tokens to represent different concepts, i.e. V_i^*, and optimize them along with pivotal W_V parameters in self-attention and W_K and W_V in cross-attention layers.

Concept Enhancement Strategy. In joint training on multiple concepts, it is essential that V_i^* for different concepts are distinct. However, our observations indicate that using single concept identifier token for each concept leads to similar V_i^*. Fig. 3b compares outcomes of using one concept identifier token versus multiple tokens for each concept. For simplicity, this discussion focuses on two concepts.

Initially, we examine the cosine similarity of two learned concept identifier tokens after processing through text encoder when only V_1^* and V_2^* are utilized as text prompt for generation (Similarity T in Fig. 3b). This approach results in a similarity exceeding 99%, rendering it challenging to differentiate these two concepts under these conditions. To address this limitation, we augment input text prompts with more muscial description, changing it to 'V_1^*, *Description*' and 'V_2^*, *Description*' (Similarity T + P in Fig. 3b). This modification reduces the similarity, but it is still above 60%.

These similarity scores are indicative of the discriminative capacity of the concept identifier tokens, a crucial factor for generating optimal music that incorporates multiple concepts. When the similarity score is high, V_1^* and V_2^* are likely to converge on the same concept, leading the model to generate music that predominantly reflects one concept while neglecting the other. The ΔAudio Alignment Score (details can be found in Evaluation Metrics.) further substantiates this, showing a significant discrepancy in Audio Alignment Scores between the two concepts when only a single concept identifier token is used for each concept. Higher ΔAudio Alignment indicates the model is more likely to generate only one concept rather than the simultaneous generation of the two concepts as we expect.

Based on this experiment, we increase concept identifier tokens number for each concept. This concept enhancement strategy significantly improves model's discriminative ability for multiple concepts, ensuring a more accurate representation in complex musical compositions. Applying the strategy leads to a reduction in all key metrics presented in Fig. 3b. This decline in metrics indicates the enhanced discriminative ability of our model when handling multiple concepts.

Experiment

In this paper, we establish a new benchmark, which includes both Dataset and Evaluation Protocol, to facilitate the customized music generation task. We also show implementation details of our experimental approach. Then, we present a comparative analysis of our method against a selection of baseline models to highlight its efficacy. Finally, the paper concludes with an in-depth ablation study, providing insights into contributory elements of our method.

Dataset and Evaluation

Dataset. We collected a benchmark of 20 distinct concepts, including a balanced collection of 10 musical instruments and 10 genres. For each concept, we collect a two-minute audio segment to form the training set. We also collected 20 prompts from MusicCap (Manco et al. 2021) dataset, which

	Tuned Parameters	Text Alignment ↑	Audio Alignment ↑	Preference Ratio ↑
Training Concept Token Only	0.001M	34.70	27.41	5.3
Training All Parameters in U-Net	746.02M	15.89	61.65	7.8
Training All Parameters in U-Net & Concept Token	746.03M	12.31	63.37	4.3
Training Cross-Attn KV & Concept Token	25.56M	26.60	23.30	10.5
Training additional LoRA Parameters	6.63M	20.32	44.40	21.8
JEN-1 StyleDreamer-Single	26.18M	29.39	37.07	50.3
JEN-1 StyleDreamer-Multiple	26.81M	22.24	44.73	–

Table 1: Quantitative comparisons. Our method achieves the best two-type alignment balance.

were utilized to evaluate the versatility and robustness across various musical themes. In our evaluation suite, we generated 50 audio clips for each concept and prompt. We will make both the dataset and evaluation protocol public available, to facilitate future research in subject-driven audio generation. More details are shown in Supplementary Materials.

Evaluation Metrics. We evaluate our method based on three metrics, the first two of which are similar to those proposed in Textual Inversion (Gal et al. 2022).

(A) **Audio Alignment Score**, which measures the similarity between the generated audio and the target concept. It shows the model's ability to learn new concepts from the reference music. Specifically, the CLAP (Elizalde et al. 2023) model is utilized to calculate the CLAP space features. The cosine similarity between features from the generated audio and the target concept is calculated to determine the Audio Alignment Score. In the context of multi-concept generation, the audio alignment for each target concept within the generated audio is computed separately. The mean of these values is then taken as the final Audio Alignment Score.

(B) **Text Alignment Score**, which evaluates the ability of methods to generate target concepts that are aligned with corresponding textual prompts. After generation, we computed the average CLAP-space feature of these generated audios. The Text Alignment Score is then determined by calculating the cosine similarity between this average CLAP-space feature and the CLAP-space features of the textual prompts without the concept identifier token V^*.

(C) **ΔAudio Alignment score**, which is utilized only in the context of multiple-concept learning, to evaluate the model tendency. In the multiple-concept learning, the ΔAudio Alignment score is the discrepancy between the Audio Alignment Score for each target concept. Higher ΔAudio Alignment indicates the model is more likely to generate only one concept rather than the simultaneous generation of the two concepts as we expect. Our ultimate objective is to distinctly learn different concepts for multiple concepts. Therefore, a model achieving a lower ΔAudio Alignment score is considered more effective in this regard.

Implement details

We utilize JEN-1 (Li et al. 2024) model as the pretrained model. The textual condition features are extracted by FLAN-T5 (Chung et al. 2022) before sending into the U-Net model. All experiments are conducted using an A6000 GPU and Pytorch framework. More training details are shown in the Supplementary Materials.

Comparisons with baseline

We generate five baseline models for comparative analysis. As demonstrated in Table 1, ours outperforms these baselines considering the balance of Text and Audio Alignment.

Training Concept Token Only. The first baseline optimizes solely learnable concept identifier tokens for new concepts, consistent with in (Gal et al. 2022). Our approach's superiority over the first baseline can be attributed to the training of a broader variety of parameters, enhancing the model's ability to extract new concepts from the reference music. In contrast, training that focuses solely on concept identifier token proves insufficient for learning concepts from reference music. While such training might yield a higher Text Alignment Score, it often results in generated music that scarcely reflects the concept of the reference. This discrepancy leads to suboptimal results in the Audio Alignment Score.

Training All Parameters in U-Net. The second baseline model diverges by keeping the tokens for new concepts fixed while fine-tuning all parameters in the diffusion model. Here, each target concept is represented by a unique identifier, such as 'sks', an infrequently used token that remains unchanged during fine-tuning, as in (Ruiz et al. 2023). While the second model trains more parameters than ours, it still underperforms, illustrating that the generation ability of a model depends not only on the quantity but also on the type of trained parameters. Specifically, training all parameters in the U-Net model can lead to substantial overfitting to the reference music, making the text prompt losing the ability to control the generation. As shown in Table 1, Training All Parameters in U-Net gets a low score in Text Alignment.

Training All Parameters in U-Net & Concept Token. In the third baseline, we optimizes the learnable concept identifier tokens as well as all parameters in the diffusion model. With more parameters to be tuned, the third model get a lower Text Alignment, showing serious overfitting issue.

Training Cross-Attn KV & Concept Token. For the fourth baseline, we limit fine-tuning the key and value projection parameters in the cross-attention layers of the U-Net, introducing a new V^* token for the new concept while keeping other parameters fixed, as in (Kumari et al. 2023). Although

Training Ratio (%)	Text Alignment ↑	Audio Alignment ↑	Preference Ratio ↑
1	29.39	37.06	17.5
5	**26.01**	**39.91**	**34.5**
10	24.11	42.23	30.3
50	19.43	46.10	8.3
100	18.67	46.68	9.4
5-random	28.14	35.64	-

Table 2: Ablation study on Training Parameter Ratio and Parameter Selection.

	Concept Tokens Number		
	1	3	5
Text-Single ↑	25.87	26.17	26.01
Audio-Single ↑	38.24	37.33	39.91
Text-Multiple ↑	21.99	22.25	17.63
Audio-Multiple ↑	42.55	44.73	44.43
ΔAudio Alignment ↓	24.38	8.05	12.20

Table 3: Ablation study on Concept Identifier Token Number for single and multiple concepts. ΔAudio Alignment is the difference between the Audio Alignment Score of two concepts for multiple-concept learning.

it incorporates learnable concept identifier tokens and partial network parameter training, falls short of our model's performance. Training only KV in cross-attention layers is not enough to learn the concept from the reference music, leading to poor performance on Audio Alignment. This highlights the necessity of carefully balancing the number of trainable parameters to effectively learn new concepts without losing the prior knowledge of the pretrained model.

Training additional LoRA Parameters. Additionally, we adopt a parameter-efficient fine-tuning (PEFT) method, LoRA (Hu et al. 2021), as the fifth baseline. Our method achieves higher scores across all metrics, demonstrating superior performance relative to other PEFT methods.

For qualitative evaluations, we employ a Preference Ratio derived from human evaluations to assess the quality of customized generation. We collect unique samples from each method and let the raters choose their preferred samples. More details are shown in the Supplementary Materials. A higher Preference Ratio indicates a stronger preference for a particular method. The results, presented in Table 1, demonstrate a significant preference for our method.

Ablation Studies

In this section, we conduct experiments to understand how different components affect the performance of our model. We focus on Pivotal Parameters selection and examine three key areas, the ratio of training parameters, training parameters selection and comparison with random selection. For integration of multiple concepts, we also investigate the effect of using different numbers of concept identifier tokens.

Training Parameter Ratio. In the Pivotal Parameters Tuning approach, we selectively train a subset of influential value projection parameters from the self-attention layers. The selection ratio is varied from 1% to 100%, as detailed in Table 2. Increasing the ratio will improve the Audio Alignment ability but hurt the generalization ability of our model. We also conducted a user preference experiment as in the table, indicating that 5% is the most preferred. At this ratio, the model effectively balances the acquisition of new concepts with the preservation of previously learned knowledge.

Training Parameter selection. In attention layers, query and key features create the attention map, which weights value features. The weighted sum of value features produces the output. For ablation, we trained W_K (W_K and W_Q can be treated as equal in self-attention layers) instead of W_V, keeping other settings unchanged. The Text Alignment and Audio Alignment were 32.38 and 31.73, respectively, highlighting the superior fitting ability of W_V.

Compared with Random Selection. Our study also includes a comparison between our Pivotal Parameters and random selection. As shown in Table 2, the comparison between '5' and '5-random' shows that training parameters chosen through our Pivotal Parameters method brings the model superior fitting capabilities and results in a better Audio Alignment compared to training random selection.

Concept Identifier Token Number. In Table 3, we present the model's performance in terms of text and audio alignment with varying numbers of concept identifier tokens. In the context of Single Concept learning, variations in the number of concept identifier Tokens show minimal impact. However, in multiple-concept learning (we use two concepts here), despite similar Text and Audio Alignment when using either 1 or 3 concept identifier tokens, the ΔAudio Alignment of using 1 concept identifier token is much higher than that of using 3 concept identifier tokens. This suggests a strong bias toward one of the concepts, which is contrary to our expectations for multiple-concept learning. Consequently, we have opted for using 3 concept identifier tokens in our approach to ensure a balance between distinct concept learning and computational efficiency.

Conclusion

In this paper, we introduce a new framework for customized music generation task. We utilize learnable concept identifier tokens to represent new concepts and fine-tune the large-scale text-to-music diffusion model using just a two-minute reference track. To balance the trade-off between learning new concepts while maintaining prior knowledge, we introduce a Pivotal Parameters Tuning method and optimize only the selected parameters in the diffusion model. To address the conflicting issues when introducing multiple concepts during music generation, we present a concept enhancement strategy, which greatly improves the quality of generated music featuring multiple concepts. Furthermore, we have established a benchmark and developed evaluation protocols for this customized music generation task. We anticipate that this benchmark will facilitate future research on this topic.

Ethical Statement

During the development of JEN-1 DreamStyler, we strictly followed the established ethical guidelines. The primary goal of JEN-1 DreamStyler is to inspire human creativity, fostering collaborative human-AI interactions to advance artistic innovation. We remain committed to the responsible use of AI, unequivocally opposing its use in unauthorized reproduction or plagiarism. Our approach emphasizes respect for creators' rights and advocates for the ethical integration of AI within the creative process.

References

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

Celma Herrada, Ò.; et al. 2009. *Music recommendation and discovery in the long tail*.

Chen, K.; Wu, Y.; Liu, H.; Nezhurina, M.; Berg-Kirkpatrick, T.; and Dubnov, S. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1206–1210. IEEE.

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

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and Controllable Music Generation. *arXiv preprint arXiv:2306.05284*.

Elizalde, B.; Deshmukh, S.; Al Ismail, M.; and Wang, H. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Evans, Z.; Carr, C.; Taylor, J.; Hawley, S. H.; and Pons, J. 2024a. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*.

Evans, Z.; Parker, J. D.; Carr, C.; Zukowski, Z.; Taylor, J.; and Pons, J. 2024b. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*.

Forsgren, S.; and Martiros, H. 2022. Riffusion-Stable diffusion for real-time music generation. *URL https://riffusion.com/about*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Lee, J.; Cho, K.; and Kiela, D. 2019. Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*.

Li, P. P.; Chen, B.; Yao, Y.; Wang, Y.; Wang, A.; and Wang, A. 2024. JEN-1: Text-Guided Universal Music Generation with Omnidirectional Diffusion Models. In *2024 IEEE Conference on Artificial Intelligence*.

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

Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2024. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Manco, I.; Benetos, E.; Quinton, E.; and Fazekas, G. 2021. Muscaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks*, 1–8. IEEE.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Plitsis, M.; Kouzelis, T.; Paraskevopoulos, G.; Katsouros, V.; and Panagakis, Y. 2024. Investigating personalization methods in text to music generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1081–1085. IEEE.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 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 18*, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Schneider, F.; Kamal, O.; Jin, Z.; and Schölkopf, B. 2023. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*.

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

Yao, Y.; Li, P.; Chen, B.; and Wang, A. 2023. JEN-1 Composer: A Unified Framework for High-Fidelity Multi-Track Music Generation. *arXiv preprint arXiv:2310.19180*.

Ziv, A.; Gat, I.; Lan, G. L.; Remez, T.; Kreuk, F.; Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2024. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*.