Dance-to-Music Generation with Encoder-based Textual Inversion

SIFEI LI, MAIS, Institute of Automation, Chinese Academy of Sciences, China and School of Artificial Intelligence, University of Chinese Academy of Sciences, China
WEIMING DONG*, MAIS, Institute of Automation, Chinese Academy of Sciences, China and School of Artificial Intelligence, University of Chinese Academy of Sciences, China
YUXIN ZHANG, MAIS, Institute of Automation, Chinese Academy of Sciences, China and School of Artificial Intelligence, University of Chinese Academy of Sciences, China
FAN TANG, University of Chinese Academy of Sciences, China
CHONGYANG MA, Kuaishou Technology, China
OLIVER DEUSSEN, University of Konstanz, Germany
TONG-YEE LEE, National Cheng-Kung University, Taiwan
CHANGSHENG XU, MAIS, Institute of Automation, Chinese Academy of Sciences, China and School of Artificial Intelligence, University of Chinese Academy of Sciences, China

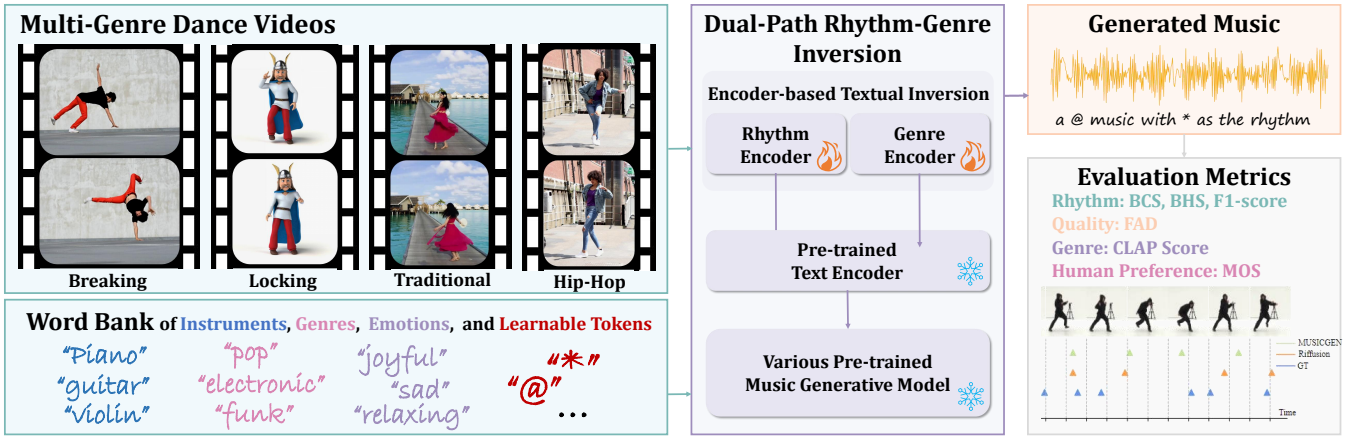

Fig. 1. We propose dual-path rhythm-genre inversion to incorporate rhythm and genre information from a dance motion sequence into two learnable tokens, which are then used to enhance the pre-trained text-to-music models with visual control. Through encoder-based textual inversion, our method offers a plug-and-play solution for text-to-music generation models, enabling seamless integration of visual cues. The word bank comprises learnable pseudo-words (represented as "@" for genre and "*" for rhythm) and descriptive music terms such as instruments, genres, and emotions. By combining these pseudo-words and descriptive music terms, our method allows for the generation of a diverse range of music that is synchronized with the dance.

*Corresponding author: Weiming Dong (weiming.dong@ia.ac.cn).

Authors' Contact Information: Sifei Li, MAIS, Institute of Automation, Chinese Academy of Sciences, China and School of Artificial Intelligence, University of Chinese Academy of Sciences, China, lisifei2022@ia.ac.cn; Weiming Dong, MAIS, Institute of Automation, Chinese Academy of Sciences, China and School of Artificial Intelligence, University of Chinese Academy of Sciences, China, weiming.dong@ia.ac.cn; Yuxin Zhang, MAIS, Institute of Automation, Chinese Academy of Sciences, China and School of Artificial Intelligence, University of Chinese Academy of Sciences, China, zhangyuxin2020@ia.ac.cn; Fan Tang, University of Chinese Academy of Sciences, China, tfan.108@gmail.com; Chongyang Ma, Kuaishou Technology, China, chongyangm@gmail.com; Oliver Deussen, University of Konstanz, Germany, oliver.deussen@uni-konstanz.de; Tong-Yee Lee, National Cheng-Kung University, Taiwan, tonylee@ncku.edu.tw; Changsheng Xu, MAIS, Institute of Automation, Chinese Academy of Sciences, China and School of Artificial Intelligence, University of Chinese Academy of Sciences, China, csxu@nlpr.ia.ac.cn.

The seamless integration of music with dance movements is essential for communicating the artistic intent of a dance piece. This alignment also significantly improves the immersive quality of gaming experiences and animation productions. Although there has been remarkable advancement in creating high-fidelity music from textual descriptions, current methodologies mainly focus on modulating overall characteristics such as genre and emotional tone. They often overlook the nuanced management of temporal rhythm, which is indispensable in crafting music for dance, since it intricately aligns the musical beats with the dancers' movements. Recognizing this gap, we propose an encoder-based textual inversion technique to augment text-to-music models with visual control, facilitating personalized music generation. Specifically, we develop dual-path rhythm-genre inversion to effectively integrate the rhythm and genre of a dance motion sequence into the textual space of a text-to-music model. Contrary to

SA Conference Papers '24, December 03–06, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1131-2/24/12
https://doi.org/10.1145/3680528.3687562

traditional textual inversion methods, which directly update text embeddings to reconstruct a single target object, our approach utilizes separate rhythm and genre encoders to obtain text embeddings for two pseudo-words, adapting to the varying rhythms and genres. We collect a new dataset called In-the-wild Dance Videos (InDV) and demonstrate that our approach outperforms state-of-the-art methods across multiple evaluation metrics. Furthermore, our method is able to adapt to changes in tempo and effectively integrates with the inherent text-guided generation capability of the pre-trained model. Our source code and demo videos are available at https://github.com/lsfhuihuiff/Dance-to-music_Siggraph_Asia_2024.

CCS Concepts: • **Computing methodologies** → **Learning latent representations**; • **Applied computing** → **Sound and music computing**.

Additional Key Words and Phrases: Dance-to-music generation; Textual inversion; Diffusion models; Pre-trained music generative models.

ACM Reference Format:
Sifei Li, Weiming Dong, Yuxin Zhang, Fan Tang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu. 2024. Dance-to-Music Generation with Encoder-based Textual Inversion. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 03–06, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3680528.3687562

1 Introducion

If life is a sort of dance, then music is the enchanting melody that accompanies our every graceful movement. Music's power lies in its ability to synchronize and enhance visual narratives, a trait extensively utilized in the gaming and film industries. The advent of short video content has seen a surge in amateur bloggers sharing dance videos on social media, showcasing music's integral role in this domain. Classical approaches [Davis and Agrawala 2018; Lee et al. 2005] align audio and video by time-warping the video. However, this process often necessitates the manual selection of suitable music and introduces variations to the original video. Additionally, copyright issues related to music can impede the widespread distribution of videos. Consequently, creating an original soundtrack for a video remains a challenging task.

As AI-generated content continues to advance, there has been a surge in research focusing on dance-to-music generation, aligning with contemporary trends and reflecting the broader evolution of AI applications in creative domains. Some approaches [Aggarwal and Parikh 2021; Di et al. 2021; Gan et al. 2020; Han et al. 2024; Su et al. 2021] employ symbolic music representation, such as MIDI and REMI, to generate music for dance videos. However, the generated music is often limited to a few classical instruments (*e.g.*, piano, guitar). Furthermore, these methods often result in incoherent note sequences and simplistic melodies, deviating significantly from the characteristics of real dance music. Other approaches [Tan et al. 2023; Yu et al. 2023; Zhu et al. 2022a,b] directly use waveforms or mel-spectrograms to generate music for dance videos, resulting in more intricate and nuanced musical compositions. However, the generated audio quality is often suboptimal, and some tend to overemphasize percussive elements such as drum beats, failing to capture melodic nuances adequately. Moreover, these approaches are inherently limited by the musical genres prevalent in the training dataset, making it challenging to generate diverse and suitable music for dance videos across various styles. Consequently, their real-world applicability remains constrained. Pre-trained text-to-music models [Copet et al. 2023; Forsgren and Martiros 2022] demonstrate their ability to generate high-quality and diverse music. However, most current models only provide control over global attributes of the music and cannot manipulate local properties such as rhythm.

To address the above-mentioned problems, we propose a novel approach for dance-to-music generation that leverages arbitrary pre-trained text-to-music models. Our method seamlessly integrates rhythm information extracted from dance videos into these models, while preserving their original text-based generation capabilities. Given an input dance video, the model generates synchronized music that harmoniously complements the dance movements. As illustrated in Fig. 1, our method facilitates dance-to-music generation across various genres (*e.g.*, "pop", "lock", "house"). Additionally, by leveraging textual descriptions, users can exert control over high-level musical attributes like genre and emotion. Notably, our approach's reliance on motion sequences for rhythm extraction allows for its extension to diverse physical activities (*e.g.*, "jump rope", "artistic gymnastics").

To achieve this goal, we develop a method to incorporate rhythm information into a pre-trained text-to-music model. Inspired by Textual Inversion [Gal et al. 2023a] which employs a pseudo-word to represent a specific concept through image reconstruction, we aim to learn a pseudo-word that represents rhythm information inherent in arbitrary dance videos and integrate it into the vocabulary of the text encoder during training. However, in the text encoder, the text embeddings corresponding to an individual token are fixed, while the rhythms in different videos are variable, making it challenging to create a pseudo-word for each rhythm pattern. Therefore, in contrast to prior methods [Gal et al. 2023a; Huang et al. 2023c; Mokady et al. 2023; Zhang et al. 2023b] using a fixed text embedding for each pseudo-word, we propose an encoder-based textual inversion method that incorporates an encoder branch that enables the text embedding associated with the pseudo-word to dynamically adapt to varying input conditions. As a result, our method enables the generation of music with different rhythms based on different motion sequences in the video.

To narrow the gap between motion sequences and text embeddings, we employ a hybrid approach that combines traditional feature extraction techniques, extracting rhythm sequences from the motion sequences, and projector networks to construct the rhythm encoder. Furthermore, we introduce a genre encoder to provide additional control over music generation.

Our contributions can be summarized as follows:

- We propose a novel encoder-based textual inversion method that enables a single pseudo-word to represent a class of variable attributes rather than a fixed object.
- We develop rhythm and genre encoders to achieve dual-path rhythm-genre inversion, converting rhythm and genre information into text embeddings. This enhances text-to-music models by integrating visual control. In dance-to-music generation, it provides flexibility in high-level feature control of music with text.
- We collect a challenging dance-music dataset entitled "In-the-wild Dance Videos" (InDV). Unlike AIST++, this dataset encompasses

a wide array of dance movements within individual videos and incorporates dance genres characterized by intricate rhythmic structures, such as Chinese traditional dance.
- Experimental results indicate that our method attains superior performance across various evaluation metrics. Additionally, our approach demonstrates robust adaptability to tempo variations and in-the-wild data.

2 Related Work

Audio-video synchronization. Some works [Davis and Agrawala 2018; Lee et al. 2005; Sun et al. 2023] achieve audio-video synchronization through editing techniques. Among them, [Davis and Agrawala 2018] and [Sun et al. 2023] enable alignment between audio and video and propose methods for quantifying video rhythm. Specifically, Davis and Agrawala [2018] calculate a directogram using optical flow to detect visible impacts, while Sun et al. [2023] learn visual eventfulness to capture video rhythm. While these works focus on editing methods, our approach extracts video rhythm from 2D keypoints and embeds it into a pre-trained model to generate new music that synchronizes with the videos.

Generating audios synchronized with input videos [Du et al. 2023; Jin et al. 2022; Qi et al. 2023; Su et al. 2023] has become increasingly popular in recent years. Some works [Gan et al. 2020; Su et al. 2020, 2021] use symbolic representation for video music generation. Controllable Music Transformer (CMT) [Di et al. 2021] designs rule-based rhythmic relationships, allowing control over timing, motion, beat, and music genre for background music generation. DANCE2MIDI [Han et al. 2024] constructs the D2MIDI dataset for multi-instrument MIDI and dance pairing, enabling the generation of coherent music sequences from dance videos. Video2Music [Kang et al. 2024] and V-MusProd [Zhuo et al. 2023] leverage multiple video features and transformer models to generate music that is synchronized with videos. However, symbolic representation-based methods overlook timbre and dynamics, limiting musical expressiveness and variation in the generated music.

Recent studies have investigated the use of spectrograms or audio waveforms for video-to-music generation. Dance2Music-GAN (D2M-GAN) [Zhu et al. 2022a] is an adversarial multi-modal framework that generates complex music samples conditioned on dance videos. Zhu et al. [2022b] introduce a novel approach that maximizes mutual information using a conditional discrete contrastive diffusion (CDCD) loss to generate dance music. LORIS [Yu et al. 2023] employs a latent conditional diffusion probabilistic model and context-aware conditioning encoders for synthesizing long-term conditional waveforms. Tan et al. [2023] combines a UNET-based latent diffusion model and a pre-trained VAE model to generate plausible dance music from 3D motion data and genre labels. However, most of them generate low-quality audio with noticeable noise, and the generated music tends to be similar, making it challenging to adapt to diverse video scenarios in the wild. In contrast, our approach, based on a pre-trained text-to-music model, is capable of generating diverse and high-quality beat-aligned dance music. V2Meow [Su et al. 2024] utilizes a multi-stage autoregressive model to generate visually-aligned music, allowing for high-level feature control with text. However, it requires training from scratch on a large-scale O(100k) dataset, in contrast to our approach, which employs a small-scale O(1k) dataset. By leveraging encoder-based textual inversion, we drive a pre-trained text-to-music model to achieve dance-to-music generation.

Text-to-music generation. Text-to-music generation has made remarkable advances in recent years. Any-to-Any generation such as CoDi [Tang et al. 2023] and NExT-GPT [Wu et al. 2023c] have the capability to generate across different modalities, including text-to-music generation. LLARK [Gardner et al. 2023] combines a pre-trained music generative model with a pre-trained language model for music understanding. Several approaches like Make-an-Audio [Huang et al. 2023a], AudioLDM [Liu et al. 2023], AudioLDM2 [Liu et al. 2024], and Tango [Ghosal et al. 2023] utilize diffusion models to generate text-guided audio encompassing speech, sounds, and music. However, these methods face limitations in terms of data quality and model scalability, resulting in lower-quality music generation. Some works such as Archisound [Schneider 2023], Riffusion [Forsgren and Martiros 2022], Moûsai [Schneider et al. 2023], and Noise2Music [Huang et al. 2023b] leverage diffusion models to generate high quality music from textual input. Furthermore, transformer-based frameworks like MusicLM [Agostinelli et al. 2023], MUSICGEN [Copet et al. 2023], and MuseCoco [Lu et al. 2023] encode music as discrete tokens, enabling the generation of high-quality music with rich textual description. Mustango [Melechovsky et al. 2024] enhances text-to-music models with control over harmony, rhythm, and dynamics through a text format. However, these methods are limited to the single modality of music and primarily manipulate high-level features, failing to establish a connection between visual features and music. In contrast, our method employs learnable pseudo-words to achieve low-level control, enabling the generation of music that aligns with video content.

Personalization of generative models. While text-guided content generation has achieved impressive results, relying solely on text is insufficient for precise control over the generated content, especially when targeting specific objectives. Therefore, the personalization of generative models has become a recent research focus. Personalization of generative models refers to the generation of content personalized to specific objects based on text-to-content models, leveraging prompt learning or fine-tuning techniques. DreamBooth [Ruiz et al. 2023] introduces a class-specific prior preservation loss for personalized image generation. LORA [Hu et al. 2021] proposes an efficient fine-tuning method known for achieving remarkable performance in personalized generation. EDICT [Wallace et al. 2023] proposes an exact diffusion inversion method for image editing, which requires no model training/finetuning, prompt tuning, or extra data. Gal et al. [2023a] introduce a textual inversion (TI) method that progressively updates the embeddings of text placeholders corresponding to specific object's visual features within a pre-trained text encoder. Inspried by TI, many variants [Alaluf et al. 2023; Gal et al. 2023b; Huang et al. 2023c; Li et al. 2023; Mokady et al. 2023; Voynov et al. 2023; Zhang et al. 2023a,b] achieve high-quality and more controllable personalized image generation. Li et al. [2024], Plitsis et al. [2024], Novack et al. [2024], Manor and Michaeli [2024] as well as Wu et al. [2023b] (MusicControlNet) explore personalized music generation using pre-trained models. However, none of these

methods successfully established a connection between visual features and music. In dance-to-music generation, the incorporation of rhythm information from dance videos is necessary. Thus, as a new task of TI, we propose an encoder-based textual inversion approach that facilitates personalized music generation, allowing for the generation of music that is aligned with specific rhythmic video.

3 Method

In this paper, we propose an encoder-based textual inversion technique, implemented as a separate network structure, that seamlessly integrates with text-to-music generation models. They are augmented with our dual-path rhythm-genre inversion, utilizing learnable pseudo-words that integrate rhythm and genre information to guide the dance-to-music generation. During training, the prompt remains fixed as "a @ music with * as the rhythm", where "@" represents the genre description and "*" indicates the rhythm description. We obtain text embeddings for two placeholder words using separate rhythm and genre encoders, resulting in two controllable tokens. By reconstructing the target audio, we simultaneously optimize the parameters of both encoders, facilitating their mutual enhancement. In the inference phase, flexible text descriptions can be used in addition to the fixed prompt to exert control over the high-level feature of music. The inclusion of the pseudo-word "*" representing rhythm is vital to ensure the generation of music that aligns with the dance.

3.1 Encoder-based Textual Inversion

Our objective is to leverage the generative capabilities of a pre-trained text-to-music model to generate music for dance videos. Typically, a text encoder firstly tokenizes a prompt into multiple indices, each corresponding to an embedding in the corresponding embedding lookup. A text transformer then encodes these text embeddings to serve as conditional guidance for the generative model. We employ three prominent music generation frameworks as backbones: Riffusion [Forsgren and Martiros 2022], AudioLDM [Liu et al. 2023], and MUSICGEN [Copet et al. 2023]. For Riffusion and AudioLDM, the generative models employed are the Latent Diffusion Models (LDMs) [Rombach et al. 2022], with the text encoders being the text encoder of CLIP [Radford et al. 2021] and CLAP [Wu et al. 2023a], respectively. For MUSICGEN, the text encoder utilized is the T5 [Raffel et al. 2020], while the generative model is the Transformer decoder.

A prevalent issue in most text-to-music models is their limited ability to control local attributes of music (*e.g.*, rhythm), due to the difficulty of accurately describing these attributes in natural language. This limitation poses a significant challenge when generating music that aligns with dance videos, where precise control over rhythm is essential. Textual inversion [Gal et al. 2023a], a personalization method for image generation, employs a pseudo-word (*e.g.*, "*") as a placeholder for a specific object and iteratively optimizes the text embedding associated with the pseudo-word. This process aims to integrate the textual description of the specific object into the vocabulary of the pre-trained text encoder, thereby enabling targeted editing of the specific object. The optimization objective can be defined as follows:

$$v_{i*} = \arg\min_{v} \mathbb{E}_{z,y,\epsilon,t}[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2], \tag{1}$$

where $z \sim E(x), \epsilon \sim \mathcal{N}(0,1)$, ϵ_θ denotes LDMs, c_θ denotes the text encoder of CLIP, and y denotes the prompt.

It is natural to consider using a pseudo-word to represent the rhythm of a video and guide the generation of music, *e.g.*, "lively melody with * as the rhythm and easy instrumental arrangement". However, textual inversion requires training a separate model for each object, with fixed embeddings for all pseudo-words. Yet, the rhythm of dance is highly variable, and training a model for the rhythm of each video would be costly. Therefore, we propose an encoder-based textual inversion method, which allows us to define the rhythm as a variable attribute in inversion.

To enhance the expressiveness of individual pseudo-words, we introduce an encoder branch into the text encoder. This encoder extracts features from a specific input type (*e.g.*, human poses), and maps them to the textual space. This enables a single pseudo-word with the capacity to exhibit distinct text embeddings contingent upon the input, allowing for control over the generation of diverse content. The encoder is iteratively optimized by reconstructing the target objects. For Riffusion and AudioLDM, the loss function of the encoder is defined as follows:

$$L_E = \mathbb{E}_{z,x,y,\epsilon,t}[\|\epsilon - \epsilon_\theta(z_t, t, t_\theta(x, y))\|_2^2], \tag{2}$$

where x denotes a specific type of input, t_θ denotes the expanded text encoder. For MUSICGEN, the cross-entropy loss function of the encoder is defined as follows:

$$L_{\text{ce}} = \frac{1}{K}\sum_{k=1}^{K} \text{CE}(G_\theta(t_\theta(x, y))_k, t_k), \tag{3}$$

where t denotes target tokens, x denotes a specific type of input, t_θ denotes the expanded text encoder, G_θ denotes the generative model and K denotes the index of codebook.

Utilizing a single pseudo-word enables precise control over specific attributes of the generated content (*e.g.*, rhythm). This control can seamlessly integrate with human text editing, resulting in enhanced flexibility. The encoder-based textual inversion framework is not confined to a specific task but can be applied to different areas, where the attributes are variable and challenging to describe using natural language. In this study, we demonstrate our approach through dual-path rhythm-genre inversion, employing rhythm and genre encoders to facilitate dance-to-music generation.

3.2 Rhythm Encoder

We combine traditional feature extraction with a projector. "Dance to the rhythm" serves as an important connection point for synchronizing music and video. Conversely, "drop the beat to the motion" is crucial for the task of creating music for dance videos. Dancers typically perform actions or transitions in sync with specific musical beats during dancing. The initiation and transitions of movements align with points of local maximum acceleration in kinematics. Based on LORIS [Yu et al. 2023], we calculate the first-order difference of the 2D keypoints $p(t, j, c)$ of the dancers over time to obtain the motion velocities, where j represents joints and c denotes

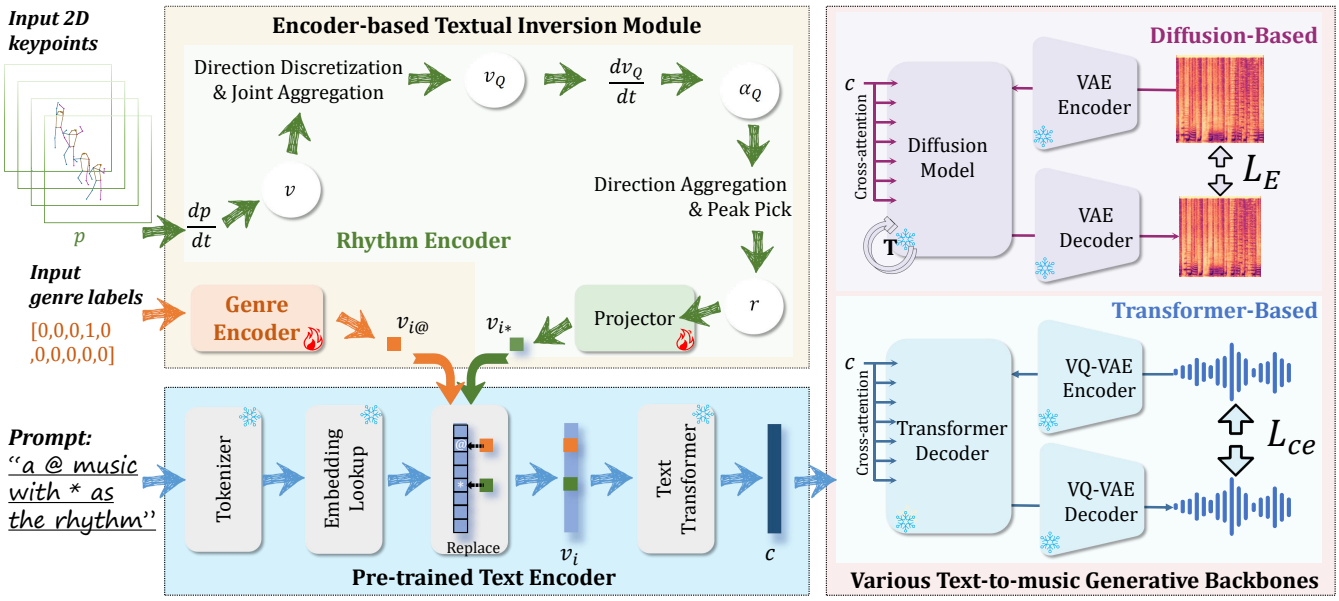

Fig. 2. We employ various pre-trained music generative models as the generative backbone and propose an encoder-based textual inversion method. During training, we fix the prompt as "a @ music with * as the rhythm", where "@" and "*" respectively represent the placeholders for the genre and rhythm of the input dance. Our dual-path rhythm-genre inversion optimizes the rhythm encoder and genre encoder together during training. Parameters v_i, $v_{i@}$, and v_{i*} correspond to the text embeddings of the prompt, "@", and "*", respectively.

coordinates. We use v_x and v_y to represent the velocities in the x and y directions, respectively. Subsequently, we discretize them into K intervals, a process which we refer to as direction discretization:

$$v_Q(t, j, k) = Vl_\theta(t, j, k), \tag{4}$$

$$l_\theta(t, j, k) = \begin{cases} 1, & \text{if } k = [\frac{\theta}{2\pi/K}] \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$V = \sqrt{{v_x}^2 + {v_y}^2}, \theta = \arctan\left(\frac{v_y}{v_x}\right), \tag{6}$$

where θ represents the direction angle, and V represents the magnitude of the velocity. By calculating the first-order difference in the temporal dimension of v_Q and aggregating the acceleration values across bins, we obtain the discrete acceleration a_Q.

We retain the positive acceleration values and compute the sum of these accelerations across joint and directional dimensions to obtain the total acceleration, denoted as a:

$$a = \sum_{j,k} a_Q(t, j, k). \tag{7}$$

Next, we determine the local maxima of a within a given time window. Subsequently, we construct a rhythm sequence r, where positions corresponding to the local maxima are set to 1, while all other positions are set to 0. We then design a projector to map r to the textual space, resulting in an embedding denoted as v_{i*}, which replaces the text embedding for the pseudo-word "*".

3.3 Genre Encoder

Reconstructing the audio mel-spectrogram solely from rhythmic information may introduce music genre information into the rhythmic pseudo-word, which hinders attribute disentanglement. Furthermore, it also results in a lack of control in music generation based on the dance genre. To overcome this, we introduce a genre encoder to enhance the generative model's reconstruction capability by incorporating genre information. We represent dance genres using one-hot encoding. By employing a combination of linear and activation layers, we map the one-hot encoding to the textual space, effectively replacing the text embedding associated with the pseudo-word "@" with the corresponding genre embedding $v_{i@}$.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate our method on two dance-music datasets, *i.e.*, AIST++ [Li et al. 2021] and our newly collected InDV dataset, both consisting of ten genres. The AIST++ dataset comprises 70 songs paired with 460 Choreographies, evenly distributed across 10 different dance genres, along with corresponding 2D keypoints. We employ the official training, validation, and testing sets and segment the data into 5.12-second clips. As a result, our final training, validation and testing sets consist of 2744, 36, and 36 samples, respectively. It is free for research purpose and we use this dataset for the main experiments and evaluations. We also collect and annotate a new dance-music dataset called InDV (In-the-wild Dance Videos), which contains 595 5.12-second clips with 216 songs. The dataset is categorized into 10 genres: Ballet, Breaking, Chinese Traditional Dance,

Table 1. Quantitative evaluation and user study results in comparison with state-of-the-arts methods on AIST++ dataset. The best results are in highlighted **bold** and the second best ones are underlined (same in the following tables).

	Rhythm				Quality	Genre	MOS		Inference Time
	BCS ↑	BHS ↑	F1-score ↑	TD ↓	FAD ↓	CLAP ↑	Coherence ↑	Quality ↑	s/clip ↓
Ground Truth	-	-	-	-	-	-	4.26	4.25	-
CMT [Di et al. 2021]	0.3368	0.1515	0.2090	21.74	16.54	0.4454	1.79	3.08	6.44
CDCD [Zhu et al. 2022b]	0.4233	0.2151	0.2852	19.25	16.47	0.3032	2.02	1.35	5.72
LORIS [Yu et al. 2023]	0.3721	0.3371	0.3537	17.80	13.15	0.6180	2.56	2.42	21.2
MDM [Tan et al. 2023]	0.3798	0.4185	0.3982	22.96	4.812	0.5793	2.97	2.58	**3.71**
Ours (AudioLDM)	0.4419	0.3605	0.3971	22.73	8.522	0.7030	2.95	**3**.02	14.82
Ours (MUSICGEN)	0.4118	0.3874	0.3992	**16.06**	6.014	0.4685	**3.56**	**3.54**	11.48
Ours (Riffusion)	**0.4761**	**0.4398**	**0.4572**	20.34	**3.416**	**0.7680**	3.24	3.15	8.49

Table 2. Comparision results of the phase-aligned version of each output. OA, OM, and OR represent Ours (AudioLDM), Ours (MUSICGEN), and Ours (Riffusion), respectively.

	Rhythm				Quality	Genre
	BCS ↑	BHS ↑	F1-score ↑	TD ↓	FAD ↓	CLAP ↑
CMT	0.4038	0.1850	0.2538	21.74	16.94	0.5135
CDCD	**0.4619**	0.2318	0.3087	21.66	16.71	0.2845
LORIS	0.4476	0.3438	0.3889	21.39	13.42	0.6216
MDM	0.3460	0.4270	0.3823	21.44	4.258	0.5889
OA	0.4073	0.3351	0.3677	26.40	8.739	0.6853
OM	0.4368	**0.4341**	**0.4354**	**15.19**	5.762	0.6384
OR	0.4452	0.4251	0.4349	20.34	**3.201**	**0.7453**

Hip-hop, Jazz, Latin, Locking, Popping, and Waacking. The training, validation, and testing sets consist of 544, 25, and 26 samples, respectively. We employ HRNet [Sun et al. 2019] in mmPose [Contributors 2020] to obtain 2D skeletons (dance motion sequences) from in-the-wild data. The video fps for the keypoints is 60.

Implementation details. We conduct experiments on our encoder-based textual inversion method using three prominent music generation frameworks: diffusion-based models Riffusion [Forsgren and Martiros 2022] and AudioLDM [Liu et al. 2023], and autoregressive model MUSICGEN [Copet et al. 2023]. For Riffusion, we use the default hyperparameters of LDMs and employ a base learning rate of 0.0005. The model training is executed on an NVIDIA GeForce RTX 3090 GPU, taking approximately 12 hours to complete over 50 epochs. As for MUSICGEN, we apply the AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.95$, and a weight decay of 0.01. A warm-up learning rate of 1e-4 is used for all layers during the initial 6000 training iterations. We train the network for 20 epochs on a single NVIDIA A40 GPU, taking approximately 10 hours. For AudioLDM, we keep the default hyperparameters and train it for 50 epoch on a single NVIDIA A40 GPU, taking approximately 11 hours.

4.2 Evaluation Metrics

Inspired by previous work and further analysis of the task, we develop a comprehensive evaluation protocol that incorporates multiple metrics to evaluate from the following perspectives.

Rhythm. Regarding the musical beats, a tolerance offset of one second is easily discernible to the human ear. To enhance the accuracy of evaluation, we adjust the tolerance offset from 1 second to 0.2 second. Specifically, we follow the setup of LORIS [Yu et al. 2023], where B_g denotes the number of musical beats in the generated music, B_t denotes the number of musical beats in the ground-truth music, and B_a denotes the number of aligned musical beats. We also conduct evaluation on global rhythm metrics. The evaluation metrics related to rhythm alignment are as follows:

- Beat Coverage Score (BCS): the ratio of aligned musical beats to the generated musical beats (B_a/B_g).
- Beat Hit Score (BHS): the ratio of aligned musical beats to the ground truth beats (B_a/B_t).
- F1-score: an integrated assessment of rhythm alignment.
- Tempo Difference (TD): average L1 norm of tempo difference between generated and ground truth music.

Audio quality. Frechet Audio Distance (FAD) [Kilgour et al. 2019] is widely used for evaluating audio quality by measuring the similarity between the generated audio and the ground truth. A lower FAD value indicates more close to the ground truth. We report the FAD based on VGGish audio embedding model [Hershey et al. 2017], which is pre-trained on the YouTube-8M audio dataset [Abu-El-Haija et al. 2016].

Genre similarity. We employ CLAP [Wu et al. 2023a], a pre-trained large-scale contrastive language-audio model, to compute genre similarity. The CLAP score evaluates the degree of similarity between the CLAP embeddings of the generated audios and the ground truth audios.

4.3 Quantitative Evaluation

AIST++ dataset [Li et al. 2021]. We conducted experiments on AIST++ dataset by deploying our method on three mainstream base models: the diffusion-based Riffusion [Forsgren and Martiros 2022], AudioLDM [Liu et al. 2023] and the autoregressive MUSICGEN [Copet et al. 2023]. Additionally, we compared our approach against four state-of-the-arts methods: CMT [Di et al. 2021], CDCD [Zhu et al. 2022b], LORIS [Yu et al. 2023], and MDM [Tan et al. 2023]. We re-implemented CDCD and LORIS and evaluated CMT and MDM using the models provided by the original authors. Our experimental results on three base models and their comparisons

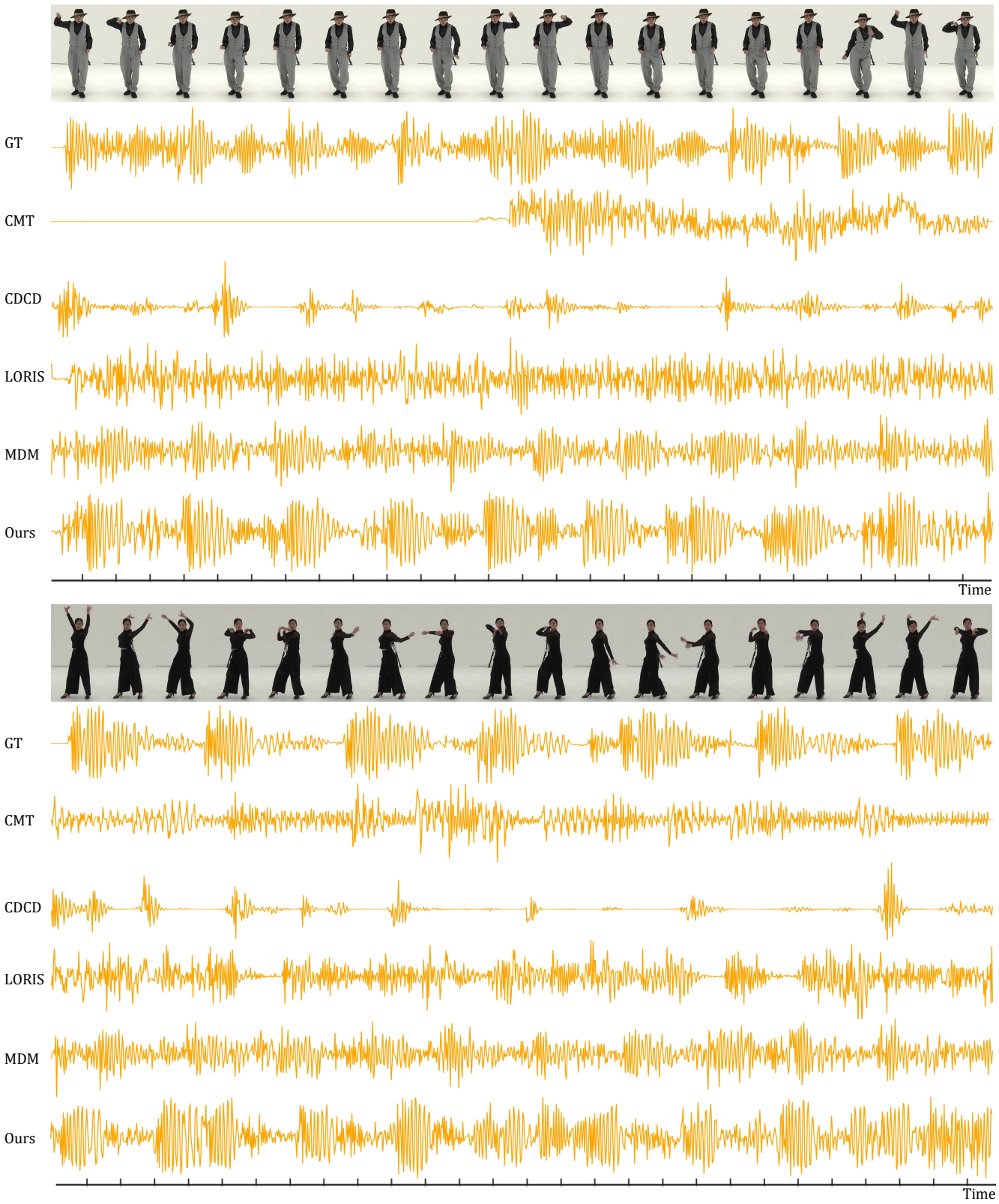

Fig. 3. Music visualization examples. “GT” and “Ours” represents “Ground Truth” and “Ours (MUSICGEN)”, respectively.

Table 3. Quantitative evaluation on InDV dataset. OM and OR represent Ours (MUSICGEN) and Ours (Riffusion), respectively.

	Rhythm				Quality	Genre
	BCS ↑	BHS ↑	F1-score ↑	TD ↓	FAD ↓	CLAP ↑
CMT	0.3513	0.2045	0.2585	19.80	12.65	0.4519
CDCD	0.3155	0.1905	0.2375	22.80	15.77	0.0747
LORIS	0.2782	0.2624	0.2701	26.15	13.51	0.2975
MDM	0.2439	0.1985	0.2189	21.98	11.30	0.4404
OM	0.3594	0.4012	0.3792	18.80	5.791	0.4048
OR	**0.3613**	**0.4562**	**0.4032**	**12.57**	**5.633**	**0.5161**

Table 4. Comparison results with music generated at the target tempo.

	Rhythm			Quality	Genre
	BCS ↑	BHS ↑	F1-score ↑	FAD ↓	CLAP ↑
Mustango	0.3849	0.3247	0.3522	13.33	0.4566
Phase-aligned Mustango	0.3964	0.3260	0.3578	14.08	0.5230
Ours (MUSICGEN)	0.4118	0.3874	0.3992	6.014	0.4685
Ours (Riffusion)	**0.4761**	**0.4398**	**0.4572**	**3.416**	**0.7680**

with other methods are shown in Table 1. Our method outperforms others across all metrics. Specifically, Riffusion-based experiments achieve optimal results on all objective metrics except TD, while MUSICGEN-based experiments achieve the best results in subjective metrics. The utilization of large-scale pre-trained models leads to longer inference times. We additionally compare the phase-aligned versions of each output (see Table 2), and our method achieves the optimal overall rhythm alignment (F1-score), indicating that our method achieves local adaptation. The slight weakness of BCS might be due to CMT, CDCD and LORIS tending to generate fewer beats. While global alignment easily improves their beat precision (BCS), the recall rate (BHS) still shows a significant weakness. In terms of genre similarity and audio quality, our method maintains an obvious advantage. The rhythm alignment metrics for certain superior methods decline because the global offset disrupts the local alignment of the generated results to some extent.

InDV dataset. Compared to AIST++, our InDV dataset is a more challenging dataset with in-the-wild video settings that contains diverse dance movements. Table 3 shows the results of the quantitative evaluation of the experiments in the InDV dataset. Our method outperforms other approaches across multiple metrics, which demonstrates the overall robustness of our approach.

Comparison with music generated at the target tempo. We use Mustango [Melechovsky et al. 2024] to generate music at the same tempo as the groundtruth and perform phase alignment. The comparison results are shown in Table 4. Our method significantly outperforms the results of Mustango, indicating that our approach provides local adaptation compared to the global control of the tempo.

User study I. We conduct a Mean Opinion Score (MOS) test. We provide participants with guidelines outlining the evaluation criteria for dance-to-music generation prior to the test. During the test, 82 participants rate a total of 80 samples, including 10 samples from each method and their corresponding ground truth. The evaluations are based on two criteria: dance-music coherence and audio quality. Dance-music coherence refers to the degree of alignment between the music and the dance video in terms of rhythm and style. Audio quality refers to the musicality of the audio (*e.g.*, noise level, richness of melody). The rating scale ranges from 1 (best) to 5 (worst). As shown in Table 1, our approach based on MUSICGEN outperforms other methods in both metrics and achieves an acceptable margin compared to the ground truth.

4.4 Qualitative Results

We visualize the waveforms and corresponding dance movements for both the ground truth and the generated music (see Fig. 3). Our results show richer dynamics that closely resemble the ground truth, while maintaining alignment with the dance movements. We also compare the beats of the generated music with those of the ground truth (see Fig. 4), demonstrating consistent generation of reliable beats with offsets below 0.2. The generated audio samples, showcasing the comparison between our method and other approaches, can be accessed on the static webpage provided within the supplementary materials. The samples show that our method generates music with superior coherence and higher quality compared to other approaches.

To verify that our method can adapt to changes in tempo, we manually change the speed of dance videos. The generated audio can still be aligned with the video rhythm while exhibiting tempo variations, as shown in Fig. 5. Detailed samples can be found in the supplementary video.

Furthermore, samples in our supplementary materials also demonstrate the ability of our method to generate plausible beat-aligned dance music on in-the-wild data (*e.g.*, film clips, virtual human dance videos, and amateur dance videos). Moreover, our method performs well in various other physical activities. The generated music aligns with the emotional atmosphere depicted in the videos, showcasing the robustness of our approach and its potential for real-world applications. Benefiting from the editability of the text-to-music model, our method demonstrates the ability to generate diverse music that aligns with the same dance videos.

4.5 Ablation Study

Genre encoder. We first investigate the role of genre encoder that engage into dance-to-music generation. We introduce a variant called CLAP that replaces genre labels with audio embeddings derived from other segments of the same audio [Wu et al. 2023a]. When the genre label is replaced with CLAP audio embeddings, experimental results indicate a decrease in all metrics except TD. This observation indicates that the highly repetitive or monotonous audios in the AIST++ dataset is not suitable for facilitating the genre encoder in mapping CLAP audio embeddings to the textual space. As a result, achieving example-based genre control becomes challenging.

Rhythm encoder. We design two projectors to map rhythm sequences into the textual space: "MLP", which utilizes a combination of multiple linear layers and activation layers as the projector, and "Attn+Pos", which incorporates position embedding and attention modules as the projector. As indicated in Table 5, experiments using the Riffusion framework demonstrate improved performance with

Fig. 4. Qualitative examples of beat alignment. The time scale interval is 0.1s. The distribution of beats show that the majority of the generated beats closely match the ground truth, with offsets below 0.2s. The Riffusion-based model demonstrates slightly better beat alignment compared to MUSICGEN-based model, consistent with quantitative metrics.

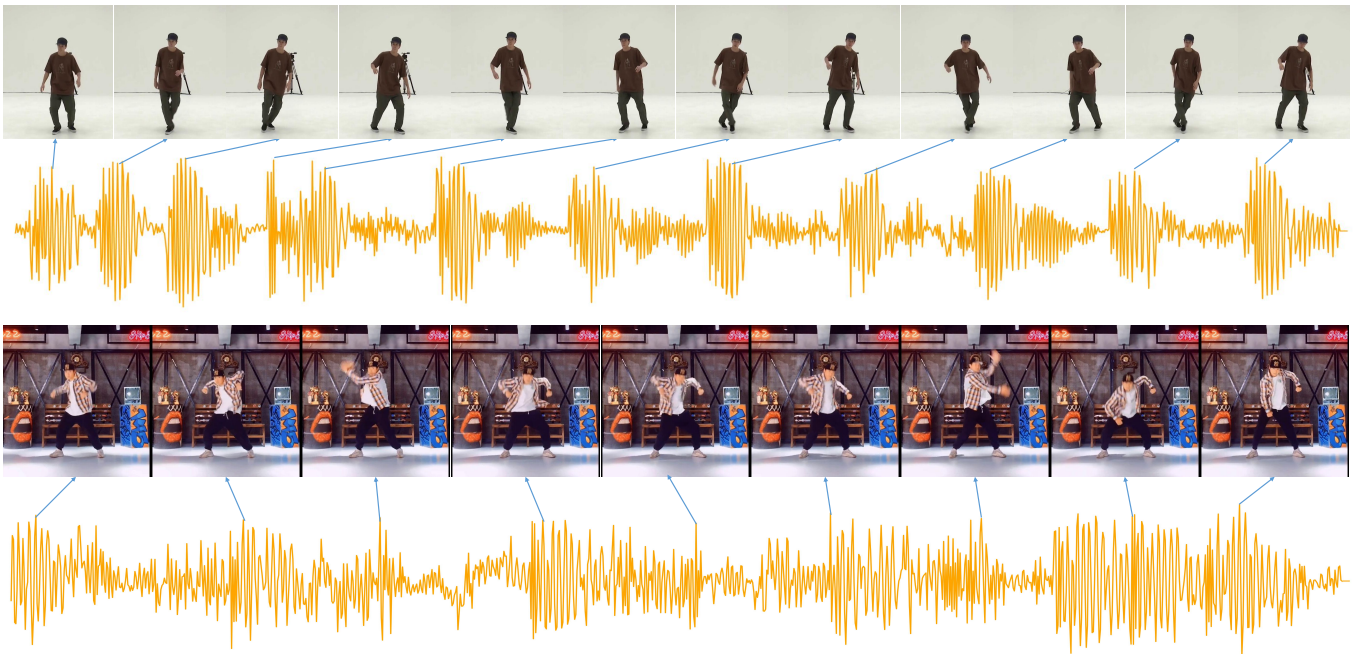

Fig. 5. Visualization of results for videos with changes in tempo. The frequency of audio peaks changes in accordance with the variations in video tempo and aligns with dance movements. The first row illustrates an example of slowed-down tempo, while the second row showcases an example of accelerated tempo.

Table 5. Ablation study of our method. "M" and "R" respectively represents "MUSICGEN" and "Riffusion". "GL" represents genre label. "CLAP" represents utilization of CLAP audio embeddings to replace genre labels. "MLP" represents the adoption of a Multi-Layer Perceptron (MLP) as the projector in the rhythm encoder. "Attn+Pos" denotes the employment of attention mechanism and positional encoding as the projector in the rhythm encoder.

Base	Encoder	Rhythm				Quality	Genre
		BCS ↑	BHS ↑	F1-score ↑	TD ↓	FAD ↓	CLAP ↑
M	CLAP	0.3925	0.3155	0.3498	**13.47**	6.821	0.4588
	MLP + GL	0.3958	0.3643	0.3794	15.51	6.469	**0.4695**
	Attn + Pos	**0.4118**	**0.3874**	**0.3992**	16.06	**6.014**	0.4685
R	CLAP	0.4480	0.3085	0.3654	**14.49**	4.611	0.6477
	MLP + GL	**0.4761**	**0.4398**	**0.4572**	20.34	**3.416**	**0.7680**
	Attn + Pos	0.3806	0.2768	0.3205	21.69	3.814	0.7524

the MLP projector but diminished results with the Attn+Pos projector. In contrast, the experiments conducted with MUSICGEN show enhanced performance with the Attn+Pos projector, suggesting that MUSICGEN is more sensitive to position embeddings.

4.6 Discussions and Limitations

Our approach generates diverse and high-quality music that aligns with videos, allowing for personalized and interactive music experiences in the context of dance videos. Furthermore, our method can be extended to in-the-wild data (*e.g.*, "film", "jump rope" and "artistic gymnastics"). For limitation, our method currently supports only fixed-length video segments for music composition. The flexibility of accommodating variable-length segments would enhance the applicability of our method in real-world scenarios.

5 Conclusion

This paper introduces an encoder-based textual inversion technique to seamlessly integrate rhythm and genre control into pre-trained text-to-music models. Our approach offers a plug-and-play solution for text-to-music models. We conduct a comprehensive evaluation on two datasets, encompassing rhythm, audio quality, and genre. Experimental results demonstrate that our method can generate high-quality music that is synchronized with the videos across various genres. Furthermore, our approach adapts to changes in tempo and enables flexible editing of high-level music attributes using text prompts. Future research may delve into the integration of multimodal motion information to achieve more robust beat generation results. Additionally, incorporating example-based genre control could provide more flexible user experiences in dance-to-music generation.

Acknowledgments

We thank Xue Song for preparing some dance video data and Minyan Luo for selection of the results. This work was supported in part by the National Natural Science Foundation of China under nos. U20B2070 and 62102162, in part by the Beijing Science and Technology Plan Project under no. Z231100005923033, in part by the National Science and Technology Council under no. 111-2221-E-006-112-MY3, Taiwan, in part by the German Research Foundation (DFG) Project under no. 508324734, and in part by Kuaishou.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv preprint arXiv:1609.08675* (2016).

Gunjan Aggarwal and Devi Parikh. 2021. Dance2Music: Automatic Dance-driven Music Generation. *arXiv preprint arXiv:2107.06252* (2021).

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. MusicLM: Generating Music From Text. *arXiv preprint arXiv:2301.11325* (2023).

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. 2023. A Neural Space-Time Representation for Text-to-Image Personalization. *ACM Transactions on Graphics* 42, 6, Article 243 (dec 2023), 10 pages.

MMPose Contributors. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and Controllable Music Generation. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*.

Abe Davis and Maneesh Agrawala. 2018. Visual rhythm and beat. *ACM Transactions on Graphics* 37, 4, Article 122 (jul 2018), 11 pages.

Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. 2021. Video background music generation with controllable music transformer. In *ACM International Conference on Multimedia*. 2037–2045.

Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. 2023. Conditional Generation of Audio from Video via Foley Analogies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2426–2436.

Seth* Forsgren and Hayk* Martiros. 2022. Riffusion - Stable diffusion for real-time music generation. (2022). https://riffusion.com/about

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023a. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *International Conference on Learning Representations (ICLR)*.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2023b. Encoder-Based Domain Tuning for Fast Personalization of Text-to-Image Models. *ACM Transactions on Graphics* 42, 4, Article 150 (jul 2023), 13 pages.

Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. 2020. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision (ECCV)*. Springer, 758–775.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. 2023. LLark: A Multimodal Foundation Model for Music. *arXiv preprint arXiv:2310.07160* (2023).

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-Audio Generation using Instruction Guided Latent Diffusion Model. In *ACM International Conference on Multimedia* (Ottawa ON, Canada). Association for Computing Machinery, New York, NY, USA, 3590–3598.

Bo Han, Yuheng Li, Yixuan Shen, Yi Ren, and Feilin Han. 2024. Dance2MIDI: Dance-driven multi-instrument music generation. *Computational Visual Media* (July 2024).

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. 2023b. Noise2Music: Text-conditioned Music Generation with Diffusion Models. *arXiv preprint arXiv:2302.03917* (2023).

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023a. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. In *International Conference on Machine Learning (ICML)*.

Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. 2023c. ReVersion: Diffusion-Based Relation Inversion from Images. *arXiv preprint arXiv:2303.13495* (2023).

Xutong Jin, Sheng Li, Guoping Wang, and Dinesh Manocha. 2022. NeuralSound: learning-based modal sound synthesis with acoustic transfer. *ACM Transactions on Graphics* 41, 4, Article 121 (2022), 15 pages.

Jaeyong Kang, Soujanya Poria, and Dorien Herremans. 2024. Video2Music: Suitable music generation from videos using an Affective Multimodal Transformer model. *Expert Systems with Applications* 249, Article 123640 (2024), 17 pages.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréechet Audio Distance: A Reference-free Metric for Evaluating Music Enhancement Algorithms. In *INTERSPEECH*. 2350–2354.

Hyun-Chul Lee, In-Kwon Lee, et al. 2005. Automatic synchronization of background music and motion in computer animation. In *Computer Graphics Forum*, Vol. 24. Amsterdam: North Holland, 1982-, 353–362.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 13381–13392.

Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. 2023. StyleDiffusion: Prompt-Embedding Inversion for Text-Based Editing. *arXiv preprint arXiv:2303.15649* (2023).

Sifei Li, Yuxin Zhang, Fan Tang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2024. Music Style Transfer with Time-Varying Inversion of Diffusion Models. In *The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 38. 547–555.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *International Conference on Machine Learning (ICML)*, Vol. 202. PMLR, 21450–21474.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. 2024. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (may 2024), 2871–2883.

Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. 2023. MuseCoco: Generating Symbolic Music from Text. *arXiv preprint arXiv:2306.00110* (2023).

Hila Manor and Tomer Michaeli. 2024. Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion. In *International Conference on Machine Learning (ICML)*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 8293–8316.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6038–6047.

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. 2024. DITTO: Diffusion Inference-Time T-Optimization for Music Generation. *International Conference on Machine Learning (ICML)*.

Manos Plitsis, Theodoros Kouzelis, Georgios Paraskevopoulos, Vassilis Katsouros, and Yannis Panagakis. 2024. Investigating Personalization Methods in Text to Music Generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1081–1085.

Qiutang Qi, Haonan Cheng, Yang Wang, Long Ye, and Shaobin Li. 2023. RD-FGFS: A Rule-Data Hybrid Framework for Fine-Grained Footstep Sound Synthesis from Visual Guidance. In *ACM International Conference on Multimedia*. 8525–8533.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 8748–8763.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22500–22510.

Flavio Schneider. 2023. ArchiSound: Audio Generation with Diffusion. *arXiv preprint arXiv:2301.13267* (2023).

Flavio Schneider, Zhijing Jin, and Bernhard Schölkopf. 2023. Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion. *arXiv preprint arXiv:2301.11757* (2023).

Kun Su, Judith Yue Li, Qingqing Huang, Dima Kuzmin, Joonseok Lee, Chris Donahue, Fei Sha, Aren Jansen, Yu Wang, Mauro Verzetti, et al. 2024. V2Meow: Meowing to the Visual Beat via Music Generation. In *The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*. 4952–4960.

Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 3325–3337.

Kun Su, Xiulong Liu, and Eli Shlizerman. 2021. How Does it Sound? Generation of Rhythmic Soundtracks for Human Movement Videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*.

Kun Su, Kaizhi Qian, Eli Shlizerman, Antonio Torralba, and Chuang Gan. 2023. Physics-Driven Diffusion Models for Impact Sound Synthesis from Videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9749–9759.

Jiatian Sun, Longxiulin Deng, Triantafyllos Afouras, Andrew Owens, and Abe Davis. 2023. Eventfulness for Interactive Video Alignment. *ACM Transactions on Graphics* 42, 4, Article 46 (jul 2023), 10 pages.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5693–5703.

Vanessa Tan, Junghyun Nam, Juhan Nam, and Junyong Noh. 2023. Motion to Dance Music Generation using Latent Diffusion Model. In *SIGGRAPH Asia 2023 Technical Communications* (Sydney, NSW, Australia) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 5, 4 pages.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-Any Generation via Composable Diffusion. *arXiv preprint arXiv:2305.11846* (2023).

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. *P+*: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522* (2023).

Bram Wallace, Akash Gokul, and Nikhil Naik. 2023. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22532–22541.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023c. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv preprint arXiv:2309.05519* (2023).

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. 2023b. Music ControlNet: Multiple Time-varying Controls for Music Generation. *arXiv preprint arXiv:2311.07069* (2023).

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

Jiashuo Yu, Yaohui Wang, Xinyuan Chen, Xiao Sun, and Yu Qiao. 2023. Long-Term Rhythmic Video Soundtracker. In *International Conference on Machine Learning (ICML)* (Honolulu, Hawaii, USA). JMLR.org, Article 1688, 15 pages.

Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023a. ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models. *ACM Transactions on Graphics* 42, 6, Article 244 (dec 2023), 14 pages.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023b. Inversion-Based Style Transfer with Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10146–10156.

Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov. 2022a. Quantized gan for complex music generation from dance videos. In *European Conference on Computer Vision (ECCV)*. Springer, 182–199.

Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. 2022b. Discrete contrastive diffusion for cross-modal music and image generation. In *International Conference on Learning Representations (ICLR)*.

Le Zhuo, Zhaokai Wang, Baisen Wang, Yue Liao, Chenxi Bao, Stanley Peng, Songhao Han, Aixi Zhang, Fei Fang, and Si Liu. 2023. Video background music generation: Dataset, method and evaluation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 15637–15647.