UniMuMo: Unified Text, Music and Motion Generation

Han Yang[1] Kun Su[2] Yutong Zhang[3] Jiaben Chen[4]
Kaizhi Qian[5] Gaowen Liu[6] Chuang Gan[4,5]
[1]The Chinese University of Hong Kong [2]University of Washington
[3]The University of British Columbia [4]UMass Amherst [5]MIT-IBM Watson AI Lab
[6]Cisco Research

Figure 1. UniMuMo is able to perform generation tasks on any combination of music, motion, and text. The tasks shown in the figure include text-to-aligned-music-motion, music-to-motion, motion-to-music, music-captioning, and motion-captioning.

Abstract

We introduce UniMuMo, a unified multimodal model capable of taking arbitrary text, music, and motion data as input conditions to generate outputs across all three modalities. To address the lack of time-synchronized data, we align unpaired music and motion data based on rhythmic patterns to leverage existing large-scale music-only and motion-only datasets. By converting music, motion, and text into token-based representation, our model bridges these modalities through a unified encoder-decoder transformer architecture. To support multiple generation tasks within a single framework, we introduce several architectural improvements. We propose encoding motion with a music codebook, mapping motion into the same feature space as music. We introduce a music-motion parallel generation scheme that unifies all music and motion generation tasks into a single transformer decoder architecture with a single training task of music-motion joint generation. Moreover, the model is designed by fine-tuning existing pre-trained single-modality models, significantly reducing computational demands. Extensive experiments demonstrate that UniMuMo achieves competitive results on all unidirectional generation benchmarks across music, motion, and text modalities. Quantitative results are available in https://hanyangclarence.github.io/unimumo_demo/.

1. Introduction

Music and body movements are synchronized and inseparable. The beat and metrical structures in rhythm encourage the spontaneous coordination of body motion with music [29], activating the motor-related areas of human brains [26]. Dance particularly exemplifies this connection through choreography that aligns with the music's rhythm, melody and emotion. Meanwhile, even though most people are not professional musicians or dancers, they often interpret music and dance using simple, natural language. This descriptive text serves as a vital bridge between understandable ideas and abstract concepts in music and motion.

The synergy between music, motion, and text provides a natural motivation to create a model capable of understanding and creating contents across all these modalities. Moreover, building a framework that can flexibly generate music, motion, and text in arbitrary combinations is crucial for real-world applications, even though existing models already achieve impressive results in unidirectional generation tasks such as text-to-music [7], music-to-motion [50], motion-to-music [55] and motion-to-text [24]. In the real world, there is a demand for diverse generative abilities, and

more complex generation tasks may be necessary, such as creating dance sequences based on both music and textual descriptions. Training individual models for each unique combination, although potentially yielding better output quality, would significantly increase training costs, deployment efforts and storage requirements. Thus, a unified model that supports all combinations of conditioning and generation tasks, rather than a collection of separate models or training adapters to incorporate individual models, offers a more cost-effective solution. To this end, we introduce a novel task of dynamically generating music, motion, and text in a multitude of combinations unifiedly. As demonstrated in Fig. 1, this task is designed to handle diverse generative scenarios, ranging from text-to-music, text-to-motion, to more complex combinations like text-to-music-plus-motion or music-plus-text-to-motion.

However, the task could be challenging, especially in two aspects: i) the lack of comprehensive datasets that include all three modalities - music, motion, and text - limits the development of a general and unified model. While there are individual datasets for music-only [44], motion-only [37], music to motion [32] and text to motion [20], a holistic and large-scale dataset that encompasses all three modalities still remains absent; ii) designing a unified architecture that supports both the conditioning and generation of all three modalities is challenging, mainly due to the significant differences between the neural representations for the three modalities and the multiplicity of desired generation tasks.

To address the first challenge of lacking paired data, we propose to align unpaired music and motion sequences based on their rhythmic patterns. Specifically, we extract both music beats and motion visual beats, then employ dynamic time warping to find the alignment and warp the motion sequence to adjust the motion visual beats to match the music beats. We found that such augmentation is accurate and efficient. With the augmented synchronized music-motion data, we can utilize existing music and motion datasets to train our unified generative model. Additionally, we construct text descriptions from music and motion metadata using a mixture of template filling, large language model generation and music-based language model generation, striking a balance between diversity, language fluency and description accuracy.

To overcome the second challenge, we propose a novel framework, UniMuMo, to unify the generation of different modalities. Our pipeline consists of three main stages: a music-motion joint tokenizer that encodes music and motion sequences into discrete representations within the same space, a music-motion transformer-decoder model trained on the task of music-motion joint generation, and a music-motion captioner that generates text descriptions from music and motion features. In the first stage, we bridge the modality gap between music and motion by mapping motion into the music feature space. Specifically, instead of using separate Vector-Quantized Variational Autoencoders (VQ-VAE) to quantize music and motion sequences, we encode motion with the codebook of a pre-trained music VQ-VAE, namely Encodec [10]. This design facilitates the unification of music and motion within the same generative framework in the subsequent stage. In the second stage, we train a unified music and motion generative model with a novel task of music-motion joint generation from text conditions. To enable the mutual conditioning of music and motion, and unlock the music-to-motion and motion-to-music generation capabilities, we introduce a novel music-motion parallel generation scheme, where we perform two mutually conditioned streams of autoregressive generation of aligned music and motion simultaneously. With the reuse of Encodec and joint encoding of motion in the previous stage, the current stage can be effectively achieved by fine-tuning the pre-trained text-to-music model associated with Encodec, namely MusicGen [7], equipping it with additional motion conditioning and generation capabilities while maintaining its music generation capabilities. In the third stage, we fine-tune a T5 decoder for music and motion captioning tasks, using the features extracted by the music-motion decoder trained in stage 2. To transform the decoder into an effective feature extractor, we replace its causal self-attention layers with trainable full self-attention layers, and fine-tune them together with the T5 decoder on music and motion captioning tasks. Extensive experiments demonstrate that UniMuMo achieves competitive performance across all unidirectional generation tasks in music, motion, and text when compared with existing state-of-the-art models, demonstrating the effectiveness and versatility of our approach.

Our work offers significant advancements in multimodal generative research, summarized as follows:

- To the best of our knowledge, this is the first unified framework capable of arbitrarily generating content across music, motion, and text.
- To address the shortage of paired multimodal data, we augment and enrich existing large-scale datasets with music-motion data alignment and text augmentations.
- We propose a novel joint codebook for encoding music and motion sequences, along with a music-motion parallel generation scheme, facilitating multiple generation tasks within a single architecture.
- Our framework achieves results comparable to SOTAs across all generation tasks in music, motion, and text.

2. Related Work

Text to Music. Text-conditioned music generation has been widely studied in recent years. There are two main branches: diffusion-based and transformer-based. For diffusion-based models, Riffusion [15] uses a latent text-to-image diffusion model to generate spectrograms, which are then con-